Critical Reading of the Scientific Literature

Dale Hancock
Field Disease Investigation Unit,
Washington State University
Pullman, WA 99164-6610

Science and technology coexist as fractious siblings in the minds of veterinarians. Technology does all the work; science merely leads us to doubt the value and truth of what we are doing. Technology, on its own, leads us into our natural habits of wizardry. Science, carried to its logical extreme, leads us into paralyzing indecision. Nowhere is this intellectual dualism more evident than in our reading, interpretation, and application of the scientific literature. On the one hand, we can find faults with the best research paper. On the other, practitioners cannot waffle endlessly but must at some point use the information at hand to make a decision.

Background: brief history of scientific thought

Rationalism: The Science of Deductive Proof.

The rationalists began with a few truths that could not be doubted (the "givens") and, from these attempted to build proofs for all of Nature in a manner similar to proving geometry theorems. Experimentation was used only to verify that which had already been deduced. Rationalism's inherent weakness soon became apparent in the telling remark of Baruch Spinoza: "...the things which I have been able to know by this knowledge so far have been very few."

Empiricism: The Science of Inductive Evidence.

David Hume turned his Scot skepticism toward the rationalists when he wrote "We are got into fairy land long ere we have reached the last steps of our theory." Like other British empiricists, Hume believed that knowledge comes only from experience. We cannot predict happenings in Nature by deduction, argued Hume; we can only induce what will likely happen based on what has happened under similar circumstances in the past.

Rationalism and Empiricism Merged.

Most scientific thought today is an mixture of rationalism and empiricism, the marriage having been commenced by the German philosopher Immanual Kant. Though schooled as a rationalist, Kant was grudgingly influenced by Hume to accept that all knowledge other than pure mathematics is based on sensory data. Deduction cannot create new knowledge. But, wrote Kant, "perceptions without conceptions are blind," meaning that our acquisition of sense data is deductively directed by innately formed, intellectual models of reality. The models can stand in for 'givens' in deductive reasoning: "Assuming that the model is essentially correct, we hypothesize that..." From these roots springs the modern Scientific Method.

Physical vs Biomedical Sciences.

The physical sciences are more rationalistic because the models tend to be highly predictive. From a model of electromagnetism we can predict the properties of a new semiconductor with some confidence. Biomedical sciences tend to be much less predictive due to large amounts of unexplained variation ("noise") that is inherent in complex systems such as the bodies of animals. Models are essential for hypothesis generation in biomedicine, but only empirical studies (e.g., efficacy trials) provide a solid basis for decision making in clinical practice.

The model and empirical science

The heart of the Scientific Method is the model and the hypotheses that we derive from it. In some disciplines, explicit model statements are included in the Methods section of scientific articles. In the veterinary literature, models tend to be fairly simple and are usually not stated explicitly. Consider the simplest model of all:

$$Y_i = U + BX_i + E$$

where

U stands for the overall population mean

Y stands for the dependent variable--the thing that we would like to have an effect on. Say morbidity in feedlot cattle or average days open in dairy cows. 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. Say a particular vaccination or a hormonal treatment for metritis.

B is the effect estimate associated with X. For example, how much does this vaccine reduce morbidity, or how much does this hormonal treatment reduce days open.

E stands for error--the unexplained noise inherent in any complex system.

The relative magnitudes of E constitute the difference between the deterministic and empirical sciences. In the physical sciences, we can usually reduce experiments to tightly controlled fragments of nature in which the unex-

plained error is small and due mainly to measurement imperfections. In medicine, we are stuck with a complex network of systems that constitute living organisms and that cannot, on an applied level, be reduced to component parts. E is almost always the largest component of a biomedical model.

Kant gave us a philosophical basis for modern science, but we can thank another German, Karl Gauss, for helping us tame error. Eliminating uncertainty is impossible, Gauss maintained, but defining it certainly is possible. From Gauss springs the statistical basis of modern empirical science.

Sources and Effects of Noise in Scientific Studies

We can reduce our interest in the critical evaluation of the scientific literature to the following questions:

1. If an article reports that X has an effect (ie, that B is non-zero), how confident can we be that this is true?

2. If an article reports that X does not have an effect, how confident can we be that it truly does not have an important effect?

3. Once we agree that X likely has an effect, how do we assess the magnitude of the effect?

4. How can we combine the results of several studies testing the same or similar X?

5. How well do the animals and—or farms in the present study relate to what is happening in our client's farms?

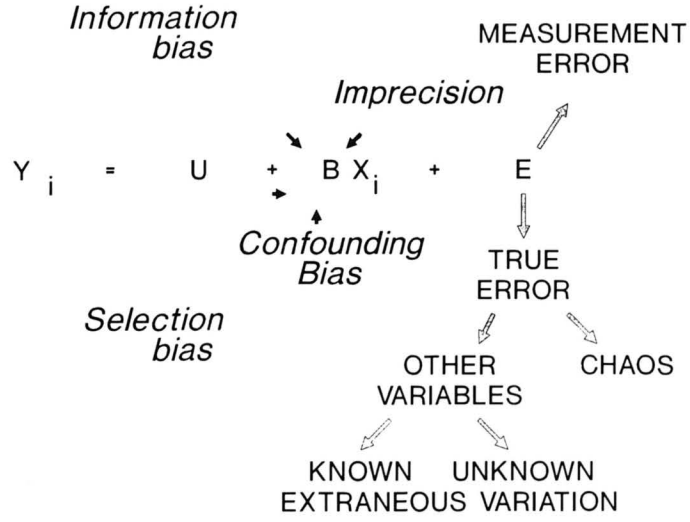
The first 3 questions relate directly to statistical error and the interactions it has with our assessment of B. Answering the 4th question requires some means of combining not only the B's from several studies but also the E's. Thus, an intuitive understanding of error is essential to critical reading of the scientific literature.

Popular views hold that error is due mainly to chaos and is therefore automatically "random error." Perhaps chaos does account for a tiny fraction of error in biomedical research, but the larger and more troublesome components derive 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 that will be discussed.

IMPRECISION AND POWER

In classical Gaussian statistics the main effect of error is imprecision. Imprecision blurs our vision so that we are unable to resolve the magnitude of B. The power of a study is the relative freedom from imprecision and thus the ability to resolve treatment effects if they exist. Power is increased primarily by increasing the number of observations.

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.



Sampling distributions.

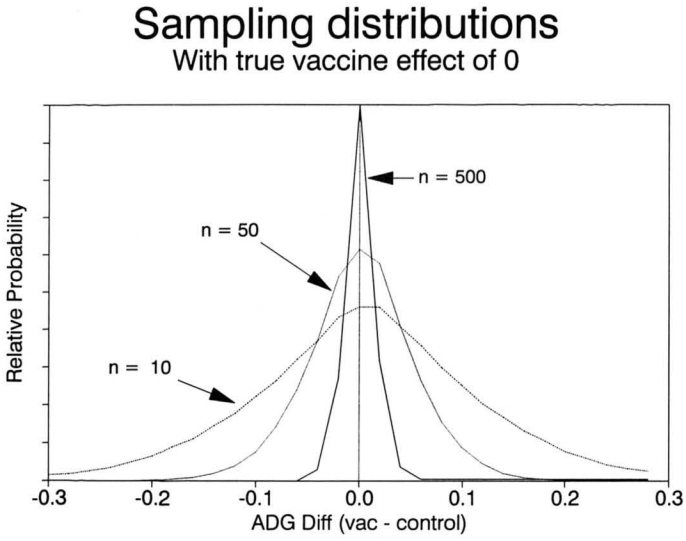
To illustrate the concept of precision, assume we want to determine the effect of a particular vaccination program on average daily gain (ADG). Just for the sake of argument, let's consider the difference in ADG between vaccinates and controls is exactly 0--that is, vaccinates and non-vaccinates have 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 (ie, because of noise). 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.

Note the 3 curves in Fig. 2. They 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 relative probabilities of observing a particular difference in ADG by 'chance' even though there is no true difference. Note that the distributions become narrower with increasing sample size; that is, 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.

Statistical significance.

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

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.



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 insufficient evidence of a favorable vaccine effect exists. In statistical parlance, we would say that the observed difference was not significant since one as large or larger than it could have occurred 'by chance' even if no true difference existed ($P > .10$). If, on the other 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.

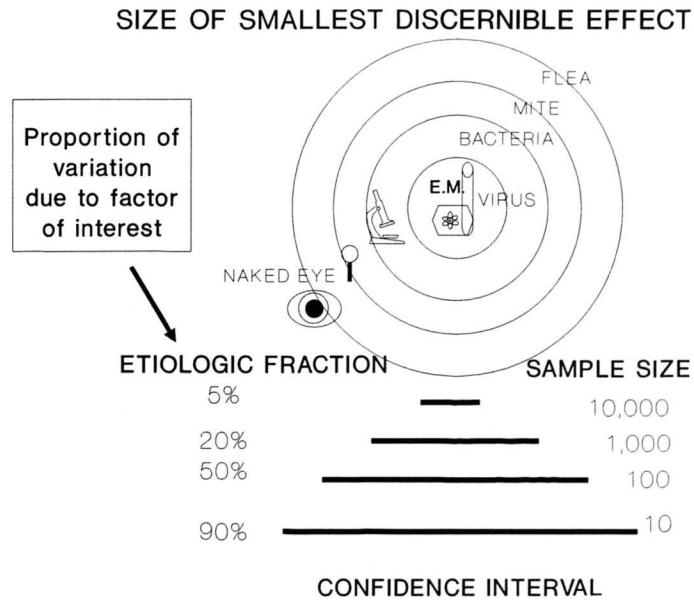
Effect "not significant".

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 but indicates that this difference was not significant. Choose one of the following interpretations:

(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 a petri dish of secretions up 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 (Fig 3). Sample size formulae, tables, and graphs are readily available for this purpose.

Figure 3. Statistical power compared to magnification power.



Ideally, scientific researchers would plan the power of their study designs, and journal reviewers would evaluate power so as to protect readers from low power studies. Until this becomes standard, we'll have to make judgements about power when we see "no significant difference" reported in a scientific paper. How do we do this? The hard way is to compute power (something we'll not cover here but for which spreadsheet templates exist). Or we could look up the sample size that would give adequate power. An indirect but much easier way will be described shortly.

Magnitude of effect (B).

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 vacci-

nates. We're 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 is 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. 4, 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.

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

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

All the computations above depend on an estimate of the standard deviation (SD). SD is a measure of noise or variation that exists in the dependent variable of interest among the individuals in the study population. It will often be reported separately for each group in a trial. To estimate the pooled standard deviation, simply average the SD's for the groups if the sample sizes are similar or take a weighted average otherwise. For studies with more than 2 groups that report analysis of variance results, the square root of the mean square error (MSE) provides an estimate of SD. If the pooled standard error (SE) is reported, assume, unless otherwise stated, that this is the SE for a single mean and not for the difference between 2 means. Obtain the SD by multiplying the square root of one of the sample sizes by the reported pooled SE and then follow computations as in Fig 4.

Is it practical for you, the reader of scientific articles, to compute confidence intervals? Yes, but even if you choose not to, you should realize that sample differences are only estimates of true differences and the degree of imprecision can be severe with even moderately large sample sizes.

Magnitude of effect vs P-values.

It is erroneous to equate significance level with importance or magnitude of effect. That is, $P < .01$ does not imply a larger or more important difference than one found significant at $P < .05$. Given sufficient sample size, the most trivial difference can be found "highly significant" ($P < .001$). Under conditions of low sample size (and thus low power), huge and important differences will usually be found "not significant" ($P > .10$). **The only good way to evaluate the magnitude of effect is through the use of confidence intervals, and importance is a biological and economic issue not directly addressed by statistical manipulations.**

Confidence intervals and "no significant difference".

You may have noted that in both of the above confidence interval examples, the intervals did not overlap 0. That is, we could be at least 95% certain that a true positive effect exists ($P < .05$). This confirmed the statistical tests. However, the reason for the confidence intervals was not merely repetitious with the statistical tests. The purpose was to decide how large or small the true effect might be. Is there a similar reason for confidence intervals where "no significant difference" is reported? Yes, this is the easy substitute for evaluating power that was mentioned earlier.

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.

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 commonly missed by authors of low power studies and 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 plausible. Most people would count a .30 lb/day increase in ADG as very important. Thus, **equating "no significant" difference with evidence of lack of an important difference would be erroneous in this example or in any low power study.**

In the indirect evaluation of the power of a study in which "no significant difference" is reported, we can compute the confidence interval and make judgements about whether the values at the extremes of the interval would be biologically (read economically) important. If either value would be considered of an important magnitude, as in the example in the previous paragraph, then we are forced to conclude that the study was not sufficiently powerful to answer the question to which it was put. Bluntly stated, after reading such a paper we still don't know didly, positive or negative, about the treatment in question.

"Had the sample sizes been larger, the differences would have been significant."

Comments such as this are usually made in the conclusions of a low power study in which 'no significant' difference was found (an outcome that is predictable at the start of such trials). The statement likely results from the authors' discomfort in negating a favored hypothesis and, perhaps, from a vague notion that, in low power studies, failure to find a significant difference does not provide evidence against the existence of a biologically important difference. Where authors of such statements err is in confusing the observed sample effect with the true (and unknown) one. In the above example, after observing a difference of .15 lb/day gain between 10 vaccinates and 10 controls, it might be tempting to conclude 'If we sample another 40 or so animals, this difference of .15 lb/day will become significant.' The trouble with this reasoning is, that if we sample another 40 animals, the observed difference is unlikely to remain .15 lb/day; indeed, based on our confidence interval it might be anywhere between -.11 lb/day and .41 lb/day.

Statistical significance and confidence intervals for occurrence data.

Much of the data reported in the veterinary literature are from occurrence phenomenon such as morbidity and mortality. All of the above comments about power and confidence intervals apply to occurrence data. The formula for the difference between 2 proportions (e.g., morbidity) and a worked example are shown in Fig 5. Sample sizes required for morbidity and mortality studies are much larger than many researchers appreciate (Fig 6). For example a feedlot vaccination trial in which controls are expected to have around 20% morbidity and it is desired to reliably detect a 25% morbidity reduction in the vaccinated group (ie, to 15%), a sample size of about 1000 animals per group would be required. Studies with substantially less sample size than suggested by Fig. 6 have low power and should be interpreted with caution.

Figure 5. Method for calculating 95% confidence interval for the difference between 2 proportions.

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

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

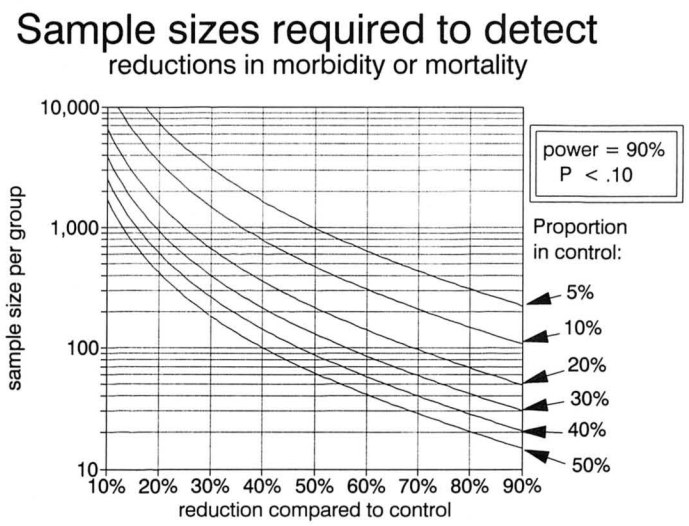
.10 +/- 2 X .086

-.07 to .27 95% confidence interval

Pens or herds as the experimental unit.

Sometimes the experimental unit is something other than individual animals. Consider, for example, a feedlot vaccination trial in which alternate lots of 200 animals are allocated to vaccine or control groups. That is, all the animals within a given lot receive the same treatment. The experimental unit is not individuals but pens, and all computations of statistical significance and confidence intervals must be computed on this basis. Analyzing the data as if individuals rather than pens randomly allocated to treatment will usually underestimate the amount of noise that truly exists. To do so is the statistical equivalent of measuring every head of wheat in 2 fields on a fertilizer trial.

Figure 6. Sample size curves for studies using occurrence data.



Thus, if it is clear that animals were assigned to treatments in groups but the analysis was performed on data from individuals, the study should be considered highly suspect.

BIAS AND INTERNAL VALIDITY

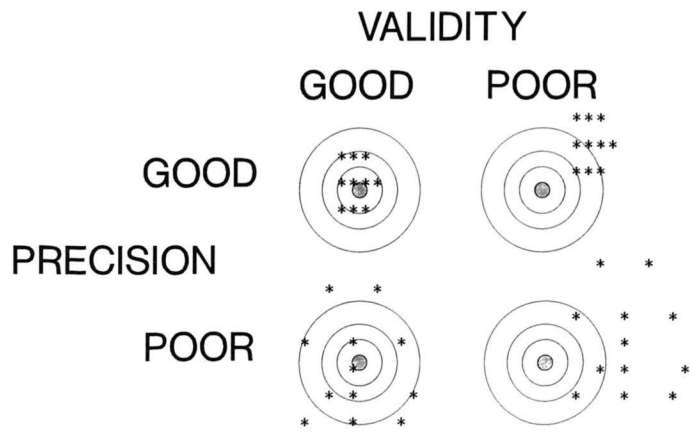
Statistical error or "noise" has 2 broad deleterious impacts on our ability to accurately estimate the effect of a treatment or management procedure. Imprecision, we have seen, reduces our ability to resolve true differences, but the problem can be avoided by providing adequate sample size. Readers of research reports can almost always evaluate the degree of imprecision and the accuracy of conclusions by, at most, a few simple computations (confidence intervals). Bias, the other deleterious impact of error, is much more insidious, and the reader must seek clues in the materials and methods as well as in the results in order to evaluate bias. The degree to which an estimate of an effect is free of bias is termed its validity.

Fig. 7 contrasts precision and validity. Where imprecision results in a random scatter of the "bullet holes" around the "bull's eye", bias results in the pattern being displaced away from the true effect. Increasing sample size will always improve an imprecise estimate but will only make a biased one worse since the pattern will "tighten" around the wrong center. Several sources of bias exist, but confounding bias is the great nemesis of empirical science and will receive most of our attention.

Confounded data.

Suppose a study was attempting to evaluate the effect of iron injections in newborn Holstein heifers on growth rate and morbidity. In the trial herd, like in most US herds, 40% of newborn calves had low passive immune levels as defined by < 5 G/dl total serum protein (TP). Suppose further that in allocating calves to treatment group, the control group ended up with 60% low TP and the treatment group with only 20%. Even if iron injections truly produce no effect on morbidity or growth rate, we would not be surprised to see an advantage (especially in morbidity) among injected calves. Our ability to evaluate the effect of iron injections is said to be confounded by TP. Even though we have every reason to believe that TP is an important variable in the health of calves, it is considered an extraneous variable here because it is not the variable of immediate interest. Confounding is produced by extraneous variables when there are unequally distributed among the treatment groups.

Figure 7. Estimation of a treatment effect compared to target shooting. Imprecision results in scatter around the bull's eye; bias results in a pattern centered away from the bull's eye.



Bull's eye = true magnitude of effect

Four ways to control confounding.

Four important tools are used in various combinations in an effort to avoid serious confounding bias.

Random allocation.

In the iron injection trial, the strong aggregation of low TP in one group should lead us to conclude that effective random allocation was probably not used. The treatment and control group will not be perfectly identical in all ways even where random allocation is used correctly, but large differences are very unlikely. The beauty of effectively used random allocation is that it will equally distribute (at least roughly) a myriad of potential confounders, known and unknown. Indeed, random allocation is the only way to control for unknown confounders, and we cannot be fully confident in the validity of any effect estimate where random allocation was not used. The goal is to have groups that are equivalent in every way, known and unknown, except for the treatment variable of interest.

At least 3 good methods of random allocation exist. The best is to use a random numbers table to assign animal identification numbers to treatment. Systematic allocation of alternating animals in some sequential order (say birth order or eartag number) is effective if it is faithfully followed. Another scheme that is sometimes used effectively is to place coded pieces of paper (poker chips, etc) into a paper bag (hat, urn, etc) with each code standing for a different treatment. As each animal becomes available, a marker is drawn for it and it is allocated to the appropriate group.

Assessing the random allocation scheme is the single most important aspect of critical reading of articles reporting the results of clinical trials and experiments. Some writers make this simple: they spell out the means of random allocation (on what basis was a particular animal assigned), and they show us the distribution of suspected confounders (say TP) in the different treatment groups. Unfortunately, many writers do not provide this information and we must look for clues:

a. If materials and methods state that random allocation was used but do not indicate how it was done, the most likely reason for the omission is that the scheme did not adhere to any conventional standard (ie, the ever-popular Haphazard Allocation Scheme was used instead).
b. If a randomization scheme is mentioned in the materials and methods that would be expected to result in roughly equal group sizes but the group sizes reported are greatly different, we are driven to suspect that at least some animals found their way into groups without being randomly allocated.
c. If the descriptive statistics of other variables are greatly dissimilar in the different groups, it is unlikely that an effective allocation scheme was followed (e.g., differences in TP distribution in the treatment groups of a neonatal calf trial).

Statisticians who have analyzed data from haphazardly "randomized" trials are extremely critical of the above omissions and discrepancies because they have seen the insidious effects of confounding and how common it is unless effective allocation is used. The results of studies in which such problems are observed should be considered suspect.

Restriction.

Two forms of restriction exist. In **complete restriction**, we limit our trial to animals that meet specified criteria. For example, the above mentioned iron injection trial in neonatal calves could be restricted to calves with TP > 5.0 g−Dl. Complete restriction is sometimes used in an attempt to improve power of morbidity/mortality studies. In this use, the trial is limited to high risk individuals; thus, a smaller sample size is required to detect an effect of specified magnitude. For example, a 50% reduction in morbidity can be detected with a smaller sample size if that in the control group is 50% compared to 5% (Fig 6). This use of complete restriction inevitably raises questions of external validity, a concept to be addressed later.

In **partial restriction** we attempt to provide for equal numbers of animals in the strata of potential confounding variables. We randomly allocate within these strata. For example, in the iron injection trial, we could determine a calf's TP status (< 5 or 5+ G/dl) and randomly allocate within each group. This would ensure equal numbers of calves in the TP strata, thus avoiding confounding. Though this approach has intuitive appeal, it is usually unnecessary in properly randomized and analyzed trials. We should not, therefore, discriminate against studies that do not use partial restriction even when obvious confounders exist.

Physical control of extraneous variation.

In certain types of scientific studies (clinical trials and observational studies, to be discussed below), the animals under observation are not maintained in a carefully controlled research facility but are dispersed in different locations, fed different diets, housed somewhat differently, and generally exposed to wide differences of environment. Where animals are randomly allocated across the different types of environments, failure to physically control the environment does not introduce confounding. Though, it will increase the error and should, in theory, reduce power, we'll see shortly that this disadvantage is illusory. Where randomization is not possible (observational studies), differences in environment introduce some major difficulties in interpretation discussed below under observational studies.

Statistical control of confounding.

Avoiding confounding does not require that treatment groups be identical with respect to potential confounders as long as we know what the confounding variables are and measure them for each individual in the trial. A toolbox of statistical methods has been developed to segregate the effect of confounders from the effect of interest (e.g., the treatment effect). An important bonus of this segregation is that, by removing the effects of selected extraneous variables, we inevitably reduce the noise and thus increase the power of the study. In the iron injection trial the segregation of TP from error would occur by extending the model:

$$Y_{ij} = x_i + Ct_{ij} + E$$

where Y_{ij} is, say, number of sick days for the jth calf within the ith treatment group, x_i is the treatment (iron injection) group, t_{ij} is the total protein of the jth calf within the ith treatment group, and C is change in Y for each increment change in TP (ie, the slope). [B is "understood" in this model since the treatment variable is categorical.] Compared to the simpler model in which the effect of TP is

contained in E, the power of model with Ct_{ij} is greater since E has been reduced (this assumes that C is non zero--that TP does have an effect on number of sick days).

Anyone who follows the scientific literature, particularly the Animal Sciences, encounters mathematical models far more complex than the simple example shown here. A full grasp requires a few graduate level statistics courses, but an intuitive understanding is needed to critically evaluate articles containing such models. Simply put, their purpose is to isolate the effects of selected confounders from error. This results in 2 benefits: (1) the distortion of confounding bias is eliminated if the treatment groups were not identical with respect to the confounders, and our estimate of treatment effect is thereby "adjusted" as if the confounders did not exist; and (2) by reducing the overall noise in the model we get more power out of a given sample size and can thus get a more precise estimate of treatment effect. Returning to the target shooting metaphor of Fig 7, statistical control of confounders helps to center our shot pattern around the bull's eye and to tighten up its distribution. What statistical models cannot do, however, is eliminate the confounding due to variables not in the model (ie, unknown extraneous variables). As we will see, this point is key in interpreting observational studies.

Types of studies

Three broad types of scientific studies exist based on the combination of the 3 above methods used to control confounding (Table 1).

Table 1. Types of studies based on methods used to control extraneous variation.

Type of study	Physical control	Randomization	Restriction	Statistical control
Observational	No	No	Sometimes	Yes
Clinical trial	No	Yes	Usually	Often
Experiment	Yes	Yes	Always	Sometimes

Experiment.

The classical experiment is carried out in a research facility in which the diet, housing, and all aspects of the environment are identical for all animals in the trial. Complete restriction is commonly used such that the study population is as homogenous as possible. Partial restriction (matching) may be used to further reduce the potential for confounding. Random allocation is always used in properly conducted experiments. Statistical control of confounding is sometimes used in experimental studies but less regularly than in the other 2 types.

Clinical trial.

The animals in a clinical trial are maintained in the "natural environment" (e.g., on farms). Commonly, animals in several or many environments will be used in a single clinical trial. Thus, physical control of the environment is not used. Complete restriction and partial restriction are commonly used. Random allocation is always in clinical trials; this is the feature that differentiates it from observational studies. Statistical control of confounding is commonly used.

Observational study.

In an observational study, physical control of the environment is not used; the study population is often widely dispersed in the natural environment. Random allocation is not possible because there is no treatment variable in the usual sense of the word "treatment." Rather than administering treatments to randomly allocated individuals, we observe individuals in different levels of the variable of primary interest (x) as they occur in nature. For example, the studies that have evaluated the effect of passive immunity on morbidity and mortality risk have used passive immune levels as they occurred naturally rather than assigning particular calves on particular farms to passive immunity groups (e.g., high and low TP). Complete and partial restriction are often used in observational studies. Statistical control of confounding is essential in observational studies and we should be skeptical of any such study that does not employ it.

Observational studies can be further divided into untargeted and targeted. Untargeted observational studies do not have a particular hypothesis but are intelligence gathering operations aimed at winnowing a large group of potential risk factors down to a small group that will be studied further. Targeted observational studies are designed to test one or a few hypotheses. This distinction is important because the results of untargeted observational studies are not intended and should not be interpreted as providing information for immediate on-farm application. We can tell untargeted observational studies because they evaluate the effects of many variables (sometimes dozens) in a single trial.

Advantages of clinical trials and observational studies over experiments.

Given the greater control over extraneous variation exercised in experiments, why should we give any credence to clinical trials and observational studies? There are 2 reasons. As we have seen, answering many of the questions faced in food animal practice requires studies with extremely large sample sizes. Funding for experiments with samples sizes in the hundreds or thousands is virtually impossible to obtain. Not that this has prevented low power experiments using ludicrously sample sizes of, say, 20 where 2000 were needed; the literature is chock full of them. Authors and journal reviewers seem oblivious to the problem, and it is left to the reader to sift through the expected "no- significant-differences were observed" (as any numerically literate person could have predicted) for any useful bits of information that might be present despite the

design flaws.

The physical control of extraneous variation that is touted as the strength of experiments is, in some respects, a weakness if our goal is an answer that can be applied in the world outside of research institutions. Although important insights can be gained by experiments if they are put to answering appropriate questions, we inevitably wonder if an effect seen in an artificially controlled environment will translate into benefits on farms. In general, if we are looking for information on factors affecting morbidity or mortality of livestock, we are not likely to find direct answers in experimental studies. Only observational studies and clinical trials offer the power and external validity that is needed.

Three tools in the scientific process.

The scientific method is how a researcher conducts a single study in an effort to answer a very specific question. The scientific process is how the scientific community uses the tools provided by the 3 types of studies to converge on utilitarian solutions to disease problems. Untargeted observational studies are used to plan targeted ones. These in turn often lead to experiments which are ideal for defining mechanisms and for the early stages of development of disease control products or procedures. Finally, clinical trials are used to determine efficacy of potential control strategies tentatively identified in experiments or targeted observational studies. The study type of most direct applicability to the practitioner is, therefore, the clinical trial; caution should be used in overextending the information provided by the other 2 types of studies.

Selection bias

Selection bias occurs when selection of individuals from which data are collected is somehow correlated to the treatment group into which they occur. One type of selection bias results from non- participation in observational studies. If a large proportion of contacted herds choose not to participate, it raises concerns about whether the relationship between the independent (x) and the dependent (Y) variables of interest is the same in the study population as in the population at large. Another type of selection bias is due to loss to followup in clinical trials. Where large numbers of individuals are withdrawn from the study, especially when the numbers are unequal in the different treatment groups, it raises concerns about whether the relationship between the treatment and the dependent variables of interest is the same in the study population (those remaining at the end of the study) as in the population at large. Observational studies and clinical trials should provide information on participation and loss to followup for readers evaluation. This is not always done and the reader can only look for clues such as unequal group sizes.

Measurement error and information bias

Perfect measurement is a myth, and if True Science requires perfect measurement then there is no such thing as true science. Unavoidable imperfections of measurement occur at all levels of research trials: dependent, independent, and confounding variables. Claims that we can measure morbidity or disease severity without error, even in whitest of ivory towers, are farcical. Claims that immune response or other physiological parameters can be measured perfectly are found, on close examination, to be philosophical tautologies (we can measure it perfectly if we decree that our method is the gold standard). In the large studies that are necessary to answer many of our questions, clerical mistakes will, sooner or later, result in measurement errors for even the most pristine of variables--say, live or dead, injected or not. The challenge of research is not to avoid measurement errors, although we should strive to minimize them. The challenge is to design studies in such a way that measurement error does not introduce information bias. Evaluating how well this was done is one of the main challenges of critical evaluation of scientific articles.

By looking back to the target shooting metaphor (Fig 7) we can visualize when measurement error will have the most adverse impact on our attempt to precisely and validly estimate the effect of some x. Measurement errors will tend to increase in the scatter around the bull's eye, but, as we have seen, increasing the sample size obviates this problem (within limits). If, however, measurement errors are not equally as likely in the different levels of x (the treatment variable) then information bias will result in a pattern that is displaced away from the bull's eye. The most likely source of information bias is failure to provide for blind evaluation. Evaluation is said to be "blind" when the evaluator is unaware of the treatment group to which study subjects are allocated. When the data have a large subjective component, as is always true in morbidity trials, blind evaluation is essential, and we should be very critical of studies not employing it (if it isn't mentioned in the M&M then you can bet it wasn't used). There are treatments that make blinding impossible since they produce visible markers. In such studies, we are hard pressed to accept a measurement method that has a large subjective component. Where blinding is not possible, only comparisons for objective data can be relied on (e.g., mortality rather than morbidity).

"The more I think, the more I doubt." Franscois Sanchez

In practicing technology, we have to decide for certain what we are going to do; there is no room for uncertainty in our actions. Empirical science is the instrument that lets us deal with the noise of nature--error we called it--and come to some conclusions that are the most likely to be true given the data at hand. The (supposed) deductive truths of

rationalism are comforting but the inductive probabilities of empirical science are more utilitarian. In the words of Bertrand Russell, "... we must either accept the inductive principle on the ground of its intrinsic evidence, or forgo all justification of our expectations about the future."

Veterinary medicine may be an empirical science, but most of us were trained as rationalists. When we begin to critically evaluate the literature, we are like tiny ships adrift in a stormy sea of uncertainty, desperately searching for some sheltered cove of undoubtable truth. The farther we go, the more certain we become that certainty is mythical and that we can only stay afloat by becoming good sailors.

Abstracts

Necrosis and sloughing of skin associated with limb cellulitis in four cows and a calf: predisposing causes, treatment and prognosis

J. A. Nguhiu-Mwangi, P. M. F. Mbithi, S. M. Mbiuki

Veterinary Record (1991) **129**, 192-195

Four cows and a calf with non-suppurative limb cellulitis were observed subsequently to suffer skin necrosis and sloughing in the affected limbs, either on or distal to the metacarpus or metatarsus. In comparison with six cows with suppurative *Corynebacterium pyogenes* limb cellulitis, topical therapy or the cases with skin necrosis and sloughing was adequate and the prognosis was good, when compared with the rigorous systemic therapy applied to the cows with suppurative cellulitis, some of which died. The skin necrosis and sloughing resulting from limb cellulitis seemed to be encouraged by the paucity of tissue between the skin and the bone, by the poor vascularity of the area, and by the causative bacteria.

Bleeding abomasal ulcers in dairy cows

U. Braun, R. Bretscher, D. Gerber

Veterinary Record (1991) **129**, 279-284

The clinical signs and changes in blood and rumen fluid, and the results of therapy are described in 35 cows suffering from bleeding abomasal ulcer. The most important pathological findings were moderate to severe anaemia with pale mucous membranes and tachycardia, dark coloured to black faeces, a disturbed general condition and anorexia. Two of the cows were slaughtered immediately. The others were treated by the transfusion of several litres of blood and the intravenous administration of a solution containing sodium chloride and glucose and other drugs such as calcium solution, vitamin K, vitamin C and metoclopramide. Two animals died in spite of the treatment and three had to be slaughtered because of the deterioration in their condition. The other 28 cows recovered within a few days and their general condition, appetite and defecation returned to normal.