Seven Ways You Can be Misled by Efficacy Studies

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

Arguably, evaluating and comparing the efficacies of products is the most important activity of bovine practitioners. Some practitioners may believe that any legally marketed products must be effective due either to government mandate or to free market forces which would drive an ineffective product to ruin. Yet, there seems little basis for either notion. Even a relatively non-critical review of literature related to extant licensed biologicals leads one to conclude that many of them are—at best—of questionable efficacy. And, given the ease with which, frankly, quack products (e.g., various "immunological" and "nutritional" supplements) are marketed over a period of many years, it is difficult to support a conclusion that the free market does a good job of "efficacy screening."

One can argue for or against the proposition that companies ought to be able to sell anything which is reasonably safe, making whatever claims they want to make, with the buyer left to sort out efficacy issues. Regardless of one's position on the moral issue, *caveat emptor* is the system which prevails; and it is, therefore, up to the practitioner to make efficacy evaluations so that he or she can provide quality service for clients.

Ideally, in evaluating efficacy of a particular product, one would have access to efficacy studies conducted by disinterested researchers—that is, those who would be equally happy to tell us that the tested product was very effective, minimally effective, or completely ineffective. Yet, such studies are comparatively rare. A large majority of efficacy trials are funded by companies wishing to market the tested products. Many in the remaining fraction are conducted by University faculty with patents on their minds. One need not invoke dishonesty to recognize that self interest colors the conduct, the analysis, and interpretation of efficacy trials and that this is contrary to the goals of the practitioner wishing to make decisions on the efficacy of products. That journals and their reviewers protect us from this tendency to favorably color product efficacies is a proposition which, in this writer's view, is very difficult to defend with the published literature.

The purpose of this report is to provide, in a hopefully humorous and satiric format, the seven principal means—intentional and unintentional, conscious and unconscious—by which reports of efficacy trials make biological and pharmaceutical products seem more effective than they really are.

1. Buzzards-and-eagles effect. "If you want to soar with the eagles, you have to pick meat with the buzzards." Traditional western aphorism.

If one is in charge of generating efficacy data for company Y's product, one is well advised to sponsor numerous efficacy trials. Even if a product is non-effective or only minimally effective, a range of estimated efficacies will be forthcoming from the multitude of studies. Some of the trials will support conclusions as to the level of "benefit" which would never be justified based on the totality. This is especially so if psuedoreplication (see below) or other allocation irregularities are used, since—even though the effects of such irregularities are neutral, on average—they greatly increase the variance of efficacy estimates and thus ensure that a few of the trials will show huge "benefits" of the product. Publication bias (the tendency of authors not to submit and journals not to publish "negative" results) and lamppost leaning (below) will take care of the rest.

2. The secret sauce of allocation irregularities. "The secret is in the sauce." Sipsy in *Fried Green Tomatoes,* describing how to cook good redneck barbecue.

Allocation is the single most important aspect of an efficacy trial, and irregularities in allocation—even seemingly minor ones—can make the desert bloom and can turn humble diluent into life-giving ambrosia. Allocation irregularities need not be intentional to work their wonders. Simple haphazard allocation will increase the variance of the efficacy estimate, even when done without bias. The secret sauce of haphazard allo-

cation combined with the buzzards-and-eagles effect (above) nearly guarantees a tasty outcome. The veterinary literature also contains examples of seemingly purposeful (not to say conscious) allocation. Examples include vaccination trials in which pre-vaccinal titers showed a substantially more naive control group and endometritis treatment studies in which subjective pre-treatment severity codes differed tremendously among treatment groups.

One can recognize haphazard or purposeful allocation irregularities by three tell-tale signs: 1) failure to describe allocation method beyond empty phrases such as "animals were randomly assigned" (a phrase commonly associated in the veterinary literature with flagrantly non-random allocation); 2) unequal sample sizes among groups, unless explicitly justified in the scheme; 3) differences among groups in terms of descriptive statistics such as age, breed, pre-treatment severity code (for therapy trials), and pre-vaccinal titer. Fortunately, users of the secret sauce need not parade their ingredients before the readers, since journal reviewers rarely seem interested in sauce recipes.

Psuedoreplication is yet another very tasty variety of secret sauce, which goes particularly well with feedlot studies. Its brewing involves randomizing at the pen level (with, say, 100 animals per pen) and then analyzing the data as if individual cattle were randomly allocated. In simulations using realistic feedlot morbidity data, repeatedly allocating 10 pens to each of two groups resulted in "significant differences" in about two-thirds of trials when the data were analyzed as if individual cattle were the allocated units (i.e., with a bogus sample size of 1000 independent individuals per group). This occurred even in the absence of any simulated treatments. For veterans of agricultural statistics courses, psuedoreplication in a feedlot is the equivalent of counting every head of wheat as an independent observation in a yield study instead of the conventional—if boring—standard of measuring bushels per acre for each plot or subplot (depending on design). The strong effect of psuedoreplication in feedlot studies is due to the skewed distribution of pen morbidities such that, even though most pens have a comparatively low morbidity, an occasional "pen crash" occurs. Stated another way, there is a strong pen (plot) effect which, by all rights, ought to be accounted for in the analysis (ideally, by using pen as the unit of observation). Fortunately, few journal reviewers know anything about psuedoreplication. While it's true that this brand of secret sauce will work against one's favored hypothesis as often as for it, combining psuedoreplication with the buzzards-and-eagles effect is an unbeatable recipe for "establishing efficacy" as long as one's product is not blatantly toxic.

3. Lamppost leaning. ***"He uses statistics as a drunken man uses lampposts—for support rather than illumination."*** **Andrew Lang, Scottish author.**

The real beauty of data is that it is subject to interpretation, and most readers can be driven to the desired interpretation like cattle to molasses. Among the most frequently used tools in this regard is the term "percent reduction" in morbidity or mortality. For example, if one observes 12% morbidity in the control group and 9% in the treatment group, this becomes a 25% reduction in morbidity. One needn't mention the huge confidence intervals around these intervals, particularly if they overlap 0 (i.e., no difference between groups). In conjunction with a healthy dollop of secret sauce (see above) or a little strategic rain dancing (see below) one can generate really impressive numbers for percent reductions in morbidity and/or mortality. From such numbers, one can go on to estimate figures like benefit:cost ratio, which will encourage readers to envision themselves counting cash on a beach somewhere. Fortunately, few readers have pondered the conundrum that, if all the "percent reduction" and "percent improved performance" claims were true, cattle producers everywhere would be living lives of indolent luxury, leaving nobody to take care of the cattle!

Another use of the lamppost effect is in "spin-doctoring" unfavorable results. Suppose one has sponsored several trials of a product. Suppose further that one of them shows a large "benefit" while the others are equivocal. Statistically literate readers will understand that a range of outcomes is expected from multiple trials—even if groups are not treated differently. Especially where allocation is by pens rather than individual cattle, "benefits" of >40% reduction in morbidity/mortality are expected to occur in 10% or so of trials even if the tested product is mere diluent. Fortunately, there are few statistically literate readers in the world; and, evidently, this is even the case for journal reviewers. Thus, one can write something like: "Even though the product was not always effective in reducing morbidity, in some trials it greatly reduced morbidity." One can then go on to discuss the "insurance benefit" and the long term benefits accruing to a large operation that—through the use of the test product—greatly reduced morbidity, if only in a small percent of pens.

A third variant of the lamppost effect involves accentuating certain features of the design in an attempt to bury deficiencies. For example, one could go to great lengths describing the efforts made to provide for blind evaluation (see item 6 below), hopefully concealing the use of, say, psuedoreplication. Or, a drawn-out description of an unnecessarily complex statistical analytical procedure can be used to bury sins of almost any variety or number. Ideally, the terminology used will be so

complex as to be beyond the comprehension of most readers, since this will lead them to skip the materials and methods section altogether (this variant is inexplicably known in some circles as "baffling them with male bovine feces").

4. Rain dancing. "Timing is everything if you're doing a rain dance." Traditional western aphorism.

Rain dancers and those who use historical controls for efficacy trials share many features, not the least of which is that the rewards realized from their activities are dependent on when they chose to begin. Given the right timing, the results can be most impressive. Notable in this regard are those efficacy trials in beef calves which use last year's calf crop as "controls." As anyone associated with the cattle business knows, unusually high rates of morbidity and mortality occur among beef calves every 4 or 5 years. One can wait patiently until a problem year comes along, then say "OK, here's the control year; we'll collect data from vaccinates next year." Using this method, one could potentially establish the appearance of "benefit" even for mild toxins.

5. Anecdote assembling. "...the plural of anecdote isn't data." Susan Dentzer, columnist for US News and World Report.

One difference between an anecdote and a datum is that the former has survived a great deal more winnowing. It's true that the buzzards-and-eagles effect (described above) does create a certain level of winnowing which is helpful to establishing a product as "beneficial" and that this does create a certain anecdote-like air for almost all commercially-sponsored efficacy trials. But for real anecdotes, the sky is the limit. Let there be ten herds on the continent out of the half million or so present which experience some presumed benefit after the use of one's product, and one has the makings of an anecdote assemblage. Another nice thing about anecdotes is that one needn't be specific. This creates the ability to say things that one *feels* are true even when one doesn't have enough data to report actual numbers: "The diarrheal morbidity, which used to be around 25%, decreased by about five-fold after the use of such-and-such product."

Though one will have difficulty in getting anecdote assemblages into the published literature in a stand-alone format, they can be included in case reports with most of the same benefits. It is interesting how journals are compelled to maintain such a high level of science for describing pathologies and microbiologies (over which one has no direct control) while the anecdotal format has become an accepted standard in the case-report literature when referring to those things (management practices) over which one does have control. Case reports, therefore, remain the best refuge for anecdotally-supported emotions and feelings about products (i.e., clinical impressions) which needn't be sullied by actual data.

On a historical note, anecdote assemblages are probably the most ancient of methods for establishing "efficacy," having been used to confirm the beneficial effect of the old standby treatment for hollow-horn: sawing off the horns and pouring kerosene into the sinuses. That some cows could be identified which lived and did very well after being so treated was irrefutable proof of efficacy, even though this once highly regarded procedure has now been replaced by more modern treatments such as megadoses of vitamin C.

6. White powder effect. "The only thing I can't figure out is what the white powder in the other vial is for." Participant in vaccine trial who had just provided an exuberant testimonial for a tested product.

The white powder effect, also known under the less colorful name of placebo effect, occurs when the person assessing outcome knows which animals belong to which treatment group. When outcome variables have a subjective component—such as respiratory or diarrheal morbidity or degree of lameness—the favored hypothesis will invariably benefit from such knowledge. This is the case when owners are collecting disease data (as in the above quote), and it also occurs when practitioners or researchers are evaluating outcomes. The fuzzier the assessment criteria the greater the placebo effect one can expect (Can you think of a non-fuzzy criterion for "diarrhea"?). Only rarely will journal reviewers make authors specify whether the outcome assessments were done blindly and will usually not reject a paper which fails to provide blind evaluation. The skeptical reader will be alert to the fact that any study employing blind assessment would likely mention it and that the remainder should be heavily discounted. Fortunately, skeptical readers are rare.

7. The Aristotle anomaly. "Aristotle could have avoided the mistake of thinking that women have fewer teeth than men by the simple device of asking Mrs. Aristotle to open her mouth." Bertrand Russell.

In reading Greek philosophers (no doubt, a common avocation among bovine practitioners) one is sometimes lead to ask: Why didn't they just look? Greek philosophers were noted for their tendency to theorize endlessly, building "new truths" from reassembled

"known truths." Today we call this tendency metaphysics as contrasted to scientific empiricism, which arose much later in Britain. The basic tenet of empiricism is that things (cows, diseases, nature, etc.) are too complicated to figure out and that the *only* basis for knowledge is to look (i.e., to collect data, as in an efficacy trial). Fortunately for those who are trying to market products, veterinary education of the mid-to-late 20th century has remained a lot closer to Greek metaphysics that to empirical science (whence the pedantical phrase, "Figure it out from what you know."). This allows someone with a few bits of agreed-upon knowledge to metaphysically construct a whole alternate universe of "new knowledge" supporting the use of a particular product: "Because high levels of such and such antigen are present in this product, it will do a better job of preventing such and such disease than will brand X." "Since use of such and such product protected those two colostrum-deprived, dexamethasone-doped calves we challenged intra-cranially with 10^9 organisms which were homologous to the vaccine strain, it will surely protect calves from natural exposure." "Including our novel adjuvant creates a better immune response and will thus improve the efficacy of such and such vaccine" (using the word "novel" will lend credibility to anyone's metaphysics). **Only the most skeptical readers will ask "Why didn't they just look?" and only the rare cynic will conclude "Maybe they did."**

Abstract

Evaluation of doramectin in a programme for season-long control of parasitic gastroenteritis in calves

M.A. Fisher, D. E. Jacobs, M.J. Hutchinson, A.J. Simon
Veterinary Record (1995) 137, 281-284

Doramectin was used in a strategic programme for the prevention of parasitic gastroenteritis in first season grazing calves. Three groups of nine calves were used: group 1 was left untreated, group 2 was treated with doramectin at 0.2 mg/kg at turnout and again eight weeks later, and group 3 was treated with 0.2 mg/kg ivermectin at three, eight and 13 weeks after turnout. Both treatment programmes prevented the gastroenteritis which occurred in the controls. The growth rates of the treated calves were superior, and their fecal egg output, and serum pepsinogen and gastrin concentrations were all substantially lower than those of the control calves. The numbers of *Ostertagia* species larvae on the pastures grazed by the treated calves were also lower than on the pastures grazed by the control calves.