CVMA Proceedings

The scientific component of this issue is devoted to the Proceedings of the Scientific Program of the 1968 CVMA Convention held in Charlottetown, P.E.I. Included are papers on clinical epidemiology, health management, and practice management, which we hope will be of interest to many readers. We apologize for the delay in publication of these proceedings caused by late submissions. The articles have been edited, but were not subjected to peer review.

The Editor

Compte-rendus de l'ACV

Les articles scientifiques de ce numéro sont des résumés des conférences présentées lors du congrès de l'ACV tenu à Charlottetown (Î.-P.-É.) en 1986. Les articles traitent de la gestion de la santé, d'épidémiologie et de l'administration d'hôpital, sujets qui, nous l'espérons, sauront intéresser plusieurs lecteurs. Vu qu'un certain nombre d'articles nous sont parvenus en retard, il ne nous a pas été possible de les publier plus tôt. Les articles n'ont pas été soumis au processus de révision.

Le rédacteur

Clinical Epidemiology

S. Wayne Martin and Brenda Bonnett

Department of Veterinary Microbiology and Immunology, Ontario Veterinary College, University of Guelph, Guelph, Ontario N1G 2W1

Abstract

Rational clinical practice requires deductive particularization of diagnostic findings, prognoses, and therapeutic responses from groups of animals (herds) to the individual animal (herd) under consideration This process utilizes concepts, skills, and methods of epidemiology, as they relate to the study of the distribution and determinants of health and disease in populations, and casts them in a clinical perspective.

We briefly outline diagnostic strategies and introduce a measure of agreement, called kappa, between clinical diagnoses. This statistic is useful not only as a measure of diagnostic accuracy, but also as a means of quantifying and understanding disagreement between diagnosticians. It is disconcerting to many, clinicians included, that given a general deficit of data on sensitivity and specificity, the level of agreement between many clinical diagnoses is only moderate at best with kappa values of 0.3 to 0.6.

Sensitivity, specificity, pretest odds, and posttest probability of disease are defined and related to the interpretation of clinical findings and ancillary diagnostic test results. An understanding of these features and how they relate to ruling-in or ruling-out a diagnosis, or minimizing diagnostic errors will greatly enhance the diagnostic accuracy of the practitioner, and reduce the frequency of clinical disagreement. The approach of running multiple tests on every patient is not only wasteful and expensive, it is unlikely to improve the ability of the clinician to establish the correct diagnosis.

We conclude with a discussion of how to decide on the best therapy, a discussion which centers on, and outlines the key features of, the well designed clinical trial. Like a diagnosis, the results from a clinical trial may not always be definitive, nonetheless it is the best available method of gleaning information about treatment efficacy.

Can Vet J 1987; 28: 318-325

In this paper we briefly introduce an epidemiological approach to some of the concepts and strategies of making a diagnosis, ways of improving diagnostic accuracy, and an introduction to the evaluation, and hence selection, of specific therapy. Many practitioners may initially find the approach and terminology confusing, if not stilted, however they have proven useful in human medicine and now constitute a significant proportion of the undergraduate epidemiology curriculum at both the Ontario Veterinary College, and the Atlantic Veterinary College. For details, the reader is referred to Clinical Epidemiology by Sackett *et al* (1) and/or Veterinary Epidemiology by Martin *et al* (2).

For the clinician, the major purpose of making a diagnosis is the classification of the patient into a suitably defined group so that the subsequent acts, be they surgical, medical, or other form of treatment, will optimize/maximize the patient's health. To the extent that the successful resolution of the illness is dependent on these acts, the diagnosis needs to be accurate (correct). Ensuring that the diagnostic process leads to the correct answer(s) is inherently difficult. First, the clinician is faced with the task of deciding which is the most likely diagnosis given that a certain set of signs and other findings is present. This contrasts with the clinical curriculum in most schools, as well as the format of most textbooks, which describe the likely distribution of signs given that a certain illness is present. Second, signs and findings are rarely pathognomonic and, in addition, the ill patient may not have a set of signs that is sufficient for the clinician to definitively establish a diagnosis. Therefore, the diagnostic process produces a "probabilistic" result. In order to increase the likelihood (posterior probability) of an accurate diagnosis, clinicians frequently will utilize one or more paraclinical tests, in addition to the findings from the history and physical examination. The correct selection and interpretation of these tests is a third complicating factor.

Although, it is not our intent, in this paper, to fully discuss *clinical diagnostic strategies*, they usually fall into one of the following: *pattern recognition*; a *multiple branching*

technique (if yes, do A; if no, do B and continue until a diagnosis is made); an *exhaustive approach* based on performing a "complete" history and physical (and paraclinical testing); or the *hypothetico-deductive process*, wherein a list of probable (as opposed to possible) differentials is established and then the list is sequentially shortened based on the history, signs, and test results present in the patient. Research and clinical testing has shown that almost all practitioners use some form of the latter approach and it is the preferred overall diagnostic strategy (1). The hypothetico-deductive process may be extended and elaborated through the use of decision analysis and decision trees.

A common feature of the diagnostic process is that the clinician observes whether or not a suggestive history is present, whether or not a sign is present, and, if paraclinical tests are used, which if any of the test results are outside of the "normal range". By combining the information from these, the clinician makes a judgment call as to the most likely diagnosis from the short list of differential diagnoses which he/she has formulated as mentioned above. The result of each aspect of the history (e.g. age, breed, sex, time since parturition), of each presenting sign (e.g. lameness, diarrhea, coughing), of each clinical finding (e.g. heart murmur, tender abdomen), and of each test (e.g. white blood cell count, liver enzyme level) serves to shape the clinician's belief that a particular disease is present or absent. In all cases this judgment is posterior; that is, the likelihood of disease is determined by the results of the finding(s). More specifically, the clinician judges the probability (albeit usually qualitatively rather than quantitatively) of a disease being present given the presence or absence of a certain finding. The following discussion is one formal way of elaborating, understanding, and improving this diagnostic process, based on the concepts of sensitivity and specificity.

In reality, patients either have the disease of interest (D+) or do not have (D−) the disease of interest, and for the sake of simplicity (and without detracting from the general validity of our discussion) we will assume a particular finding is either present (F+) or absent (F−). A finding may refer to an aspect of the history, a sign noted during the physical examination, or the result of a paraclinical test. [We will show subsequently how to utilize findings, such as enzyme levels, that are measured on a continuous scale.]

In the absence of any information about the patient, the prior probability (likelihood) of a disease is the percentage of patients that have that disease. Once the breed and age of the patient are known, the prior probability becomes the percentage of patients of that age and breed that have the disease of interest. In this manner, and viewing age and breed as findings, the posttest probability of the first step in the diagnostic process become the pretest probability in the second step, and so on. The posttest, or postfinding, probability of a specific disease given a finding is present [p(D + /F +) — this may also be read as the proportion with the disease among the test (finding) positive individuals, and is usually referred to as the predictive value of a positive test by epidemiologists] and the posttest probability of that disease given the finding is absent [p(D + /F−)] are shown in Table I. It is worthwhile to note that each of these posterior probabilities is influenced by the prior probability of disease (pD+), the ability of the finding to correctly classify animals with the disease of interest [i.e. its sensitivity; p(F+/D+)], and the ability of the finding to correctly classify animals without the disease of interest [i.e. its specificity; p(F−/D−) = 1 − p(F+/D−)] (Table II). In this example asking the breed and/or age represents a combined first step in the diagnostic process. Since the sensitivity and specificity of a finding are independent of the prior probability of disease, they are used as the best method of evaluating the ability of the finding to differentiate between animals with the disease of interest and animals with other conditions. Like the textbook description of diseases, *sensitivity is the probability of a finding being positive given the disease is present, and specificity is the probability that the finding will be negative given that the disease is absent*. The probability of the finding being positive when the disease is absent is 1 — specificity. Of course, whether or not the disease is present is not known in the diagnostic process. However, one can estimate the prior probability of a disease and utilize this knowledge when interpreting the finding, as shown subsequently.

Since a working knowledge of sensitivity and specificity is central to understanding this approach to the diagnostic process, some guidelines about how they are evaluated will be presented. To evaluate the sensitivity and specificity of a particular finding requires a biologically independent method of establishing whether or not the disease of interest is present — this is referred to as the "gold standard" by Sackett *et al* (1). Second, unless the finding is objectively measured, the presence or absence of the finding should be evaluated "blindly"

TABLE I
The Sensitivity and Specificity of a Finding and their Relationship to the Prior and Posterior Probabilities of Disease in a Clinical Setting

		State of Nature		
		D+	D−	
Finding (F)	F+	a	b	a+b
	F−	c	d	c+d
		a+c	b+d	n=a+b+c+d

D+ The disease of interest is present
D− Diseases that may be confused with D+ are present
F+ An aspect of the history, physical examination, or paraclinical test is positive
F− That aspect of the history, physical examination, or paraclinical test is negative
n The number of animals suspected to have D+
Prior Probability of Disease = p(D+) = (a+c)/n
Sensitivity of F = p(F+/D+) = a/(a+c)
1 — Sensitivity = False negative rate
Specificity of F = p(F−/D−) = d/(b+d)
1 — Specificity = False positive rate
Posterior Probability of D+ given F+ = p(D+/F+) = a/(a+b)
Posterior Probability of D+ given F− = p(D+/F−) = c/(c+d)

with regard to the gold standard. Third, the "nondiseased" group must only contain those animals with conditions that are likely to be confused with the disease of interest. Fourth, the characteristics of both the D+ and D− groups should reflect those in the source population (the potential population of patients) both in terms of the spectrum (i.e. stage, severity, body system involvement) of the disease of interest, and the makeup (e.g. age, breed, sex, range of other diseases) of the D− population. These four criteria are basic to the evaluation of a finding, and, without their fulfillment, knowing the correct inference to place on a finding is quite difficult if not impossible. Hence, *when reading about new diagnostic aids, the clinician should ascertain the extent to which these criteria have been fulfilled; if the criteria are not met, it is probably best not to waste your time pursuing the article in depth.* For an excellent recent example of how to identify the ability of selected signs to predict the status of cows with regard to abomassal ulcer, see Smith *et al* (3).

Given that the sensitivity and specificity of a finding are known, the clinician can then evaluate the usefulness of the finding in populations with different prior probabilities of disease. Since most clinicians don't actually quantitate these prior probabilities, we will use a probability of 5% to reflect the clinician's judgment that the disease of interest, although on the short list of diagnoses, is unlikely, a probability of 50% to reflect that the clinician is truly uncertain about whether the disease is present or absent, and a probability of 90% to reflect that the clinician is certain the disease of interest is present. [Probabilities are actually limited to the range 0−1, percentages (probability × 100) are used herein for simplicity; we hope this detail doesn't confuse the reader.] In general, for any specified sensitivity and specificity, the posterior probability of disease is directly related to the prior probability of disease; that is, the higher the prior probability, the greater the posterior probabilities (see Table II). More important, however, is the absolute difference between the prior and the postfinding probabilities of disease; in other words, the ability of the finding to alter the clinician's judgment about the likelihood of disease being present after noting whether the finding is present or absent. For most findings either a positive or negative result is useful when the prior probability of disease is close to 50%; otherwise a positive finding is informative — in the sense that it will alter a clinician's judgment — only if that finding has a high specificity, and a negative finding is informative only if that finding has a high sensitivity. As mentioned, *the majority of findings are most useful in the situation of greatest uncertainty*; that is, when the prior probability of disease is approximately 50%. Thus, if a finding has a low specificity and the prior probability of disease is high, it is probably better not to put the patient through a test procedure since the results of the procedure are unlikely to modify the clinician's judgment about the likelihood of the disease. In the example in Table II, at step 1 the posterior probabilities from the auscultation findings wouldn't likely change the clinician's belief about the likelihood of a congenital heart defect as much as the same finding (particularly if positive) in step 2. The sensitivity and specificity have not changed, but by noting the breed of dog and hence changing (increasing in this case) the prior probability of disease, the posterior probabilities are more informative. [This example assumes that breed is associated with the occurrence of a congenital heart defect.]

TABLE II
The Relationship between Prior and Posterior Probabilities of Disease in a Clinical Setting

Step 1 D+ = Congenital Heart Defect in Dogs
p(D+) = 1%: the prevalence of D + in all breeds in your practice
Suppose a finding (auscultation) has a sensitivity (% of dogs with abnormal ausculation result given D+ present) of 80%, and a specificity (% of dogs with normal auscultation result given a dog without D+) of 98%
The expected results after auscultating the dog are:

		State of Nature D+	D−		Probabilities
Ausculation	F+	8	20	28	0.29
	F−	2	970	972	0.002
		10	990	1000	0.01

Step 2 You note the dog is a miniature poodle
p(D+) = 5%: the prevalence of D + in miniature poodles in your practice (this is the postfinding probability of D+ when the finding is breed of dog)
The expected results after ausculating the poodle are:

		State of Nature D+	D−		Probabilities
Ausculation	F+	40	19	59	0.68
	F−	10	931	941	0.01
		50	950	1000	0.05

Data on the relationship of titer change to parainfluenza-3 (PI3) virus in cattle with bovine respiratory disease (cases) and nondiseased cattle housed with them (controls) and shown in Table III, and are presented to extend the discussion-of sensitivity/ specificity to situations where the finding is measured on a quantitiative scale. [These data could represent any diagnostic setting with the values being level of liver enzyme, white blood count, age of animale, a graded pain response to palpation, etc.] Assume for current purposes that criteria one through four for establishing sensitivity and specificity have been met. Titer change is expressed as the number of dilutiions difference between the acute and convalescent sera. Note that the average titer change is greater in cases (average = 4) than in controls (average = 2) and thus "on average" the cases are different from the controls in terms of PI3 titer. Despite this however, the distributions of titer, in these two groups, may overlap to such an extent that the titer by itself doesn't discriminate well between individual cases and controls. For instance, there is no large PI3 titer increase experienced by significantly more cases than controls. Hence a titer increase of seven dilutions or more is not sufficient to rule-in respiratory disease; the latter requires a specificity of 100%. At the other extreme, however, only 2% of cases had a titer decline, whereas 22% of

controls had a titer decline. Thus for practical purposes, any decline in titer would be sufficient to rule-out respiratory disease; the latter requires a sensitivity of 100%. In this example, such a rule gives a sensitivity of 98% and a specificity of 22%. If a titer change between these extremes is used as a critical level to declare the test positive or negative, the sensitivity and specificity will change, and the changes will be in opposite directions to each other. For example, using the usual guideline of a fourfold titer increase (≥ 2 dilutions) as biologically significant, the sensitivity is 82% and the specificity is 40%. At a critical titer change of four or more dilutions, the sensitivity is 58% and the specificity is 70%.

How does this information assist the use of this (or any other) finding in a clinical setting? If the P13 test at a critical value of ≥ 2 dilutions change were used in a population having a prior probability of respiratory disease of 30%, then p(D + /T +) = 0.37 and p(D + /T −) = 0.16; neither test result being informative because of the small absolute prior probabilities of disease (see Table IV). If the critical titer change was ≥ 4 dilutions, then p(D + /T +) = 0.45 and p(D+/T−) = 0.20; again, neither postfinding probability being particularly informative since they are not greatly different from the prior probability of disease. Obviously because of the large number of errors of classification, the PI3 titer change by itself is only of marginal value in classifying the health status of these cattle. Nonetheless, in a clinical setting, one should utilize the observed value of the finding as the critical value for establishing the sensitivity and specificity in working-up a particular case. Then, one combines these values with estimates of the prior probability of disease for a particular patient to derive the posterior probabilities. For example if the prior probability of respiratory disease in a recently arrived feedlot calf is 30%, if that calf has a titer increase of seven dilutions, (sensitivity = 5% and specificity = 97% at this level) the posttest probability of it having bovine respiratory disease is increased to 0.42 or 42 percent. As mentioned earlier such a small change in the probability of disease between the pre and posttest situations is unlikely to change the clinician's mind about the likelihood of respiratory disease and thus the titer is not informative. If the posterior probabilities are informative, subsequent acts can be modified accordingly as the clinician proceeds with the workup to establish the most likely diagnosis. As mentioned previously, *a major challenge to the clinician, particularly in selecting potentially harmful or expensive paraclinical tests, is to only use those paraclinical tests that are likely to provide information that can shape, not just confuse, the clinician's judgment.*

Every clinician is aware that in virtually all instances more than one finding is used in establishing a diagnosis, and may think that this circumvents some of the previous problems. However, regardless of the sequence of these findings, if the majority, or all findings must be positive to provide sufficient evidence of the disease of interest then the process will be relatively specific (few false positives) and the posterior probability of disease in positive individuals will usually be quite high. The sensitivity will be low however, thus many animals with the disease of interest will be missed. If a positive finding on only one or a few of the findings is deemed to be sufficient evidence of the disease of interest, the process will usually be quite sensitive (few false negatives) and the

TABLE III
The Distribution of Titer Change[a] to PI3 Virus in Cattle with (cases) and Cattle without (controls) Bovine Respiratory Disease

PI3 Titer Change	No. of Cases	% Cases	No. of Controls	% Controls
+7	3	5%	3	3%
+6	7	11%	5	5%
+5	14	21%	8	8%
+4	14[b]	21%	13	14%
+3	8	12%	11	12%
+2	8	12%	17[b]	18%
+1	5	8%	13	14%
0	5	8%	4	4%
−1			11	12%
−2	1	2%	6	6%
−3			2	2%
−4			1	1%
−5			1	1%
Total	65		95	

[a] Number of dilutions of titer change between day of arrival and 28-35 days later. The distributions (%) have been rounded so that each sums to 100%
[b] Average of respective distributions measured as number of dilutions of change

TABLE IV
The Change in Posterior Probability of Disease Depending on Choice of Critical Titer

Example 1 Critical Titer = 2 or more dilutions increase
Sensitivity = 83%, Specificity = 40%
The expected results are:

		Respiratory Disease +	−		Probabilities
Titer	T+	250	420	670	250/670 = 0.37
	T−	50	280	330	50/330 = 0.16
		300	700	700	300/1000 = 0.30

Example 2 Critical Titer = 2 or more dilutions increase
Sensitivity = 83%, Specificity = 40%
The expected results are:

		Respiratory Disease +	−		Probabilities
Titer	T+	174	210	384	174/384 = 0.45
	T−	126	490	616	126/616 = 0.20
		300	700	1000	300/1000 = 0.30

probability of disease in "negative" individuals will be quite low. The latter process tends to lead to over-diagnosis of disease and at the same time the probability of disease in "positive" individuals may not differ greatly from the prior probability. *For this reason the approach of running multiple tests on every patient is not only wasteful and expensive, it is unlikely to improve the ability of the clinician to establish the correct diagnosis.* In between these extremes, a more suitable selection and interpretation of test results may achieve a better balance between sensitivity and specificity. The severity of the disease to the individual patient, and the health risk that patient poses to other members of the source population are the major factors used when selecting the appropriate balance of sensitivity and specificity. Again, *one might think that reaching a correct diagnosis could be assured by using multiple findings. The problem which limits the ability is that in most instances the findings are correlated with each other and thus the value of the second finding after the first finding is noted, is decreased, relative to its singular value before the first finding is noted* (3).

The approach to diagnosis based on sensitivity and specificity just described is quite useful in both clinical and field situations. At the very least, it leads to a structured method of thinking about the process of establishing a diagnosis. It also points out explicitly that a clinical diagnosis is not always likely to be correct since the procedure tends to emphasize either sensitivity, or specificity, depending on the nature of the disease. A serious drawback to the utility of this approach is that often there is no gold standard that can be applied to typical patients without undue costs and/or inherent error. Thus, frequently, neither the true disease status, nor the sensitivity/specificity of particular findings is known. An example is viral diseases where the ultimate diagnosis depends on isolating the agent from the diseased tissue, a process that is expensive and specific, but often insensitive. When no practical gold standard exists, one must often use agreement — either agreement between clinicians, or agreement between tests — to provide guidance as to the most likely diagnosis. The philosophy is that if clinicians, or tests, agree then this provides increased evidence of the validity of their findings. If there is disagreement, and in the absence of other information to say which is likely correct, then neither clinician's opinion, nor test result, is of much value. Other reasons for disagreement are that the examination process (i.e. taking history, performing the physical examination, and using paraclinical tests) is flawed by a failure to standardize what constitutes a finding, coupled with the inability of the clinician to observe, or the test to provide, the same finding in the same way on different occasions, even in the same patient. The subject of agreement, and how it relates to repeatability and establishing a diagnosis will now be pursued in some detail.

The data in Table V are useful for the discussion of agreement. For present purposes we will assume that the disease of interest is feline pneumonia and that two clinicians independently examine 120 cats. The first clinician says nine have pneumonia, the second 31, and both say the same cats have pneumonia in four instances. [Before proceeding, it should be clear that since no "gold standard" data are given, the sensitivity/specificity of each clinician's diagnostic ability is unknown.] Is there evidence of significant agreement between the clinicians to suggest that both their findings are valid? We first note that the second clinician diagnoses pneumonia about 3.5 times more frequently than the first clinician; this is somewhat disconcerting and provides a clue that the agreement is not likely high. This does not indicate that clinician 2 has a higher sensitivity, he/she could have a lower specificity than clinician 1. Even if the two clinicians diagnosed the disease in the same number of patients, this would not necessarily indicate good agreement. More directly on the topic of agreement, we note that in 88 (4 + 84) of 120 or 73% of the cats there was agreement between clinicians. This seems like a good level of agreement particularly if the implied baseline level of agreement is 0%. However, the real baseline should be the chance level of agreement. From basic principles of biometry, the chance number of D+D+ agreements is $9 \times 31/120$ and the chance number of D−D− agreements is $111 \times 89/120$. Thus the chance level of agreement is $(2.3 + 82.3)/120 = 70.5\%$. [This is analoguous to flipping two coins simultaneously, one with a probability of landing heads equal to 9/120 — the first clinician's rate of diagnosis of pneumonia, the other with a probability of landing heads equal to 31/120 — the second clinician's rate of diagnosing pneumonia. The expected number of times *both* would land heads (i.e. D + D +) or *both* would land tails (i.e. D − D −) in 1200 tosses is 846 or 70.5% of the time.] Given this baseline, the observed level of agreement (P_o) exceeds the chance level (P_c) by only 2.5%. If there was perfect agreement between the tests, the difference $P_o - P_c$ could not exceed $100 - P_c$ (29.5% in this example) so this becomes the baseline for the difference between the observed and chance levels. The resulting statistic, called kappa (K), describes the proportion of the maximum achievable level of agreement that is realized, after adjusting for chance levels. The general formula for kappa is: $K = (P_o - P_c)/(100 - P_c)$

TABLE V
The Agreement between Two Clinicians in Diagnosing Pneumonia in Cats

		Clinician 2		
		D+	D−	
Clinician 1	D+	4	5	9
	D−	27	84	111
		31	89	120

% Observed agreement = P_o = (4 + 84)/120 = 73%
Chance number of D + D + agreements is (9 × 31)/120 = 2.3
Chance number of D − D − agreements is (89 × 111)/120 = 82.3
= Chance agreement = P_c = (2.3 + 82.3)/120 = 70.5%
Maximum possible % agreement beyond chance = 100% − 70.5% = 29.5%
Kappa = $(P_o - P_c)/(100 - P_c)$
= $(73 - 70.5)/(100 - 70.5)$
= 0.08

If there is no agreement beyond chance levels, K = 0; if there is perfect agreement beyond chance levels, K = 1. In this instance K = 0.08, a very low proportion of realized agreement. *For clinical purposes, K between 0.3 and 0.5 is acceptable, K between 0.5 and 0.7 is good, and K above 0.7 is excellent.* In this instance, unless one of the clinicians was recognized as an authority in diagnosing feline pneumonia and the other a neophyte (perhaps a student), the opinion of neither clinician has much value.

It turns out that these are real data, derived from two independent blind assessments of 120 fecal samples for K99 *Escherichia coli* (4). The usual culture and serology method results are represented by clinician 1, the results of a fluorescent antibody method by clinician 2. Although there are valid reasons as to why the test results should not agree, since both are used frequently in diagnostic laboratories one might have hoped for a larger kappa, particularly noting that five of nine culture- and serology-positive samples were negative on fluorescence. *Our experience is that when test results are rigorously evaluated in a blind manner, the extent of agreement, measured by kappa, is often lower than when "nonblind" methods are used.* Obviously the "blind" results are more likely to reflect the true characteristics of the finding.

A second example of assessing agreement is shown in Table VI where a clinician studied the repeatability of classification of body condition in goats (5). The study was done by examining 38 milking does in one herd and then returning and reexamining the same does eight days later. The does were unknown to the clinician and only identified by tags and hence the two examinations were conducted "blindly". The pertinent data and calculations are shown, and given the kappa of 0.7 it is obvious that the clinician was able to consistently assess the body condition of goats. Thus, without knowing the sensitivity and specificity of the clinician's classification ability, the kappa statistic provides a sound basis for deeming the data to be reliable — at least repeatable. In the absence of a gold standard, one would compare the diagnoses of two, or more, clinicians to assess validity, as noted previously.

It is disconcerting to many, clinicians included, that given a general deficit of data on sensitivity and specificity, the level of agreement between many clinical diagnoses is only moderate at best, with most kappa values between 0.3 to 0.6. It certainly suggests much room for improvement and among the suggested solutions are: 1) ensure the diagnostic environment is suited to the task (appropriate light, heat, silence, etc.); 2) seek corroboration of key findings — having "blinded" colleagues assist in this is quite useful; 3) record evidence using standardized terminology, then record inference; and 4) have an independent interpretation of paraclinical test results. The latter should be conducted "blindly" except where blindness must be broken to ensure that the correct paraclinical test is performed.

Having reached a diagnosis, it is now necessary to select the appropriate therapy. In consideration of this, it is important to recognize that "if it ain't broke, don't fix it", but if "it" requires fixing the clinician should set specific objectives for each selected therapy (e.g. to reduce a fever to the normal temperature range). Defining objectives is done both to focus the clinician's thinking about the use of a specific therapy as well as to provide an explicit benchmark for evaluating the therapy. Clinicians, it seems, primarily select a therapy based on a mixture of their (uncontrolled) clinical experience, extension of current concepts about mechanisms of disease, and advice from experts, other colleagues, or pharmaceutical representatives. Despite the apparent utility of these widely used procedures, much of which is debatable, *none of these methods is as good as reliance on a well-designed randomized controlled trial for guidance in selecting a therapeutic regimen.* The great failing of the previous methods is that they have a very low ability (power) to expose erroneous conclusions about efficacy, even if the observations made are accurate.

At this point, it is well to remind ourselves that although a specific therapy must either be efficacious, or not, the clinician does not know this ultimate truth. Rather, he/she should use the results of clinical trials (more preferable than the previous methods), to attempt to determine the truth about a given therapy's efficacy. As shown in Table VII only two of four possible combinations of state of nature and conclusions based on trial results are correct; these are represented by confidence level and power. *Usually, clinicians are more willing to have the trial result lead to the conclusion that a therapy doesn't work when it does (Type II Error) than have the trial results indicate that a therapy works when it doesn't (Type I Error).* In this regard, trials conducted on too few animals are not of much value, because it is predictable that the result will not lead to the rejection of the null hypothesis (i.e. acceptance that the treatment is effective) even when it should; that is the therapy will be judged to be of no value, incorrectly. Sometimes, if the treatment effect is small, or highly variable, very large trials will be required to reduce the level of Type II error to a reasonable level, say 20%. The point is that although the randomized clinical trial is the best single method of assessing a therapy's efficacy, it is not a perfect procedure. To minimize the likelihood of the trial producing results which could lead to erroneous conclusions, the design and performance of the trial must be at a high level.

TABLE VI
Agreement beyond Chance on Body Condition Scores of Milking Does in a Commercial Dairy Goat Herd in Ontario, 1984

		Examination 1			
		Thin	Normal	Fat	
Examination 2	Thin	1	0	0	1
	Normal	2	21	1	24
	Fat	0	3	10	13
		3	24	11	38

P_o = (1 + 21 + 10)/38 = 0.842 or 84.2%
P_c = (.079 + 15.16 + 3.76)/38 = 0.50 or 50%
K = (84.2 − 50)/(100 − 50)
= 0.684

TABLE VII
The Relationship between the True State of Therapeutic Efficacy and the Findings of a Clinical Trial

		True State of Nature	
		Therapy Doesn't Work	Therapy Works
Clinical Trial Results Indicate	Therapy Doesn't Work	Confidence Level (Usually 95%)	Type II Error (Usually 5-20%)
	Therapy Works	Significance Level or Type I Error (Usually 5%)	Power (Usually 80-95%)
	Sum of Probabilities	100	100

A brief summary of relevant clinical trial design features should assist the clinician to interpret the literature on therapeutic/clinical trials, and hence guide the choice of therapy. In order for the results of a trial to be of value to the clinician the trial should involve procedures that are feasible in practice, and the study patients should be similar to those seen in a clinician's practice. As a starting point, the diagnostic workup of the study subjects should be sufficiently rigorous to provide confidence that they had the disease of concern — bearing in mind that application of the gold standard may not be practical.

To assess the validity of a trial, the following considerations should be noted; our experience is that if these criteria are not stated explicitly, the criteria have probably not been fulfilled. In return, failing to fulfill some or all of these criteria doesn't automatically mean the article is useless, but it certainly greatly reduces the value of the report. The criteria are:

1) The patients should be assigned to a concurrent treatment group using a *formal random* procedure; we include systematic allocation in the latter. Any other manner of assigning study subjects, particularly using historical controls, detracts from the value of the study. As mentioned, considerable thought should be given to determining the required sample size. In domestic animals, an aggregate of animals (e.g. a litter, pen, or herd) may be assigned to treatments. If so, the subsequent analysis of results should reflect this fact. Failure to recognize this is common in the veterinary literature, and greatly reduces the value of the experiment.
2) Once allocated, the treated and control patients should be managed and followed with a similar degree of rigor. One method of ensuring this is to keep the patient's owner/manager and the clinician blinded as to the treatment status. A placebo is usually necessary to accomplish this.
3) At the termination of the trial, the outcomes for all patients who were originally placed in the trial should be included, and all clinically relevant outcomes should be reported. Thus although the therapeutic trial might focus on control of colitis, any other illnesses should be reported. Where possible one should also report the final health status of all patients on the trial, as well as outcomes in those who did and did not comply with the treatment regime. To avoid bias, those who assess the outcome(s) should be blind to treatment status wherever possible.
4) The analysis of results should involve a statistical method that is consistent with the design of the trial and should be appropriate for the type of data collected. Often, simple analyses (chi-square and t-test) will suffice, however in larger (e.g. multiclinic) studies a more complex method may be required (e.g. to control clinic to clinic variation in outcome). Because researchers are more willing to report positive than negative results, the literature is biased. This situation is made worse if many small studies are being conducted on a specific therapy, since there is a tendency to report only the "favorable" trial results.

Like a diagnosis, the results from a clinical trial may not always be definitive, nonetheless it is the best available method of gleaning information about treatment efficacy. As one example of an apparent conflict in results, two well-designed trials were conducted on the efficacy of routine gonadotrophin releasing hormone (GnRH) administration in postpartum dairy cows. In one trial, conducted in a number of herds, GnRH given at day 8-12 postpartum to cows with retained placenta produced no overall benefit (6). In a subset of "early bred" cows it appeared to improve reproductive performance. Although the experiment was well designed, the fact that the effect was only seen in "early bred" cows constrains the extrapolation of results from this trial. In the other trial, performed in one large herd, GnRH given at days 15-16 to all postpartum cows appeared to reduce reproductive efficiency (7). No benefit was seen in "early bred" cows in this trial. That the second trial was done in only one herd may limit extrapolation of results to other populations; however its rigorous performance lends credibility to the findings. Despite their disagreement, both trials have contributed greatly to our understanding of how this drug may produce its effect(s). The fact that cows in different herds *may* respond differently to the same drug should lead to an investigation of why such an interaction exists, not to a decrying of the lack of the ultimate answer from the clinical trial. As the health problems which veterinarians are asked to correct are very complex, it should not be surprising that different trials can produce apparently conflicting results. This however reinforces the need for formal well-designed clinical trials.

It is our hope that the topics discussed in this paper will help clinicians to better understand the diagnostic process, will aid the selection of efficacious therapeutic regimes, and lead to "more science" being incorporated into the art of veterinary medicine.

References

1. SACKETT DL, HAYNES RB, TUGWELL P. Clinical epidemiology: a basic science

for clinical medicine. Toronto: Little Brown and Company, 1985.
2. MARTIN SW, MEEK AH, WILLEBERG P. Veterinary epidemiology: principles and methods. Ames, Iowa: Iowa State University Press, 1987.
3. SMITH DF, MUNSON L, ERB HN. Predictive values for clinical signs of abomasal ulcer disease in dairy cattle. Prev Vet Med 1986; 3: 573-580.
4. WALTNER-TOEWS D, MARTIN SW, MEEK AH. An epidemiological study of selected calf pathogens on Holstein dairy farms in southwestern Ontario. Can J Vet Res 1986; 50: 307-313.
5. KUERSTEN KE. The prevalence of paratuberculosis and caprine arthritis-encephalitis in Ontario dairy goats. MSc Thesis, University of Guelph, Guelph, Ontario, Canada, 1985.
6. LESLIE KE, DOIG PA, BOSU WTK, CURTIS RA, MARTIN SW. Effects of gonadotropin releasing hormone on reproductive performance of dairy cows with retained placenta. Can J Comp Med 1984; 4: 354-359.
7. ETHERINGTON WG, BOSU WTK, MARTIN SW, COTE JF, DOIG PA, LESLIE KE. Reproductive performance in dairy cows following postpartum treatment with gonadotrophin releasing hormone and/or prostaglandin: a field trial. Can J Comp Med 1984; 48: 245-250.