ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Theriogenology 66 (2006) 534–542

Theriogenology

www.journals.elsevierhealth.com/periodicals/the

Critical evaluation of scientific articles and other sources of information: An introduction to evidence-based veterinary medicine

J.P. Kastelic *

Agriculture and Agri-Food Canada, Lethbridge Research Centre, Box 3000, Lethbridge, AB, Canada T1J 4B1

Abstract

The purpose of this paper is to briefly review key concepts regarding critical reading of the scientific literature to make informed decisions, in the context of evidence-based veterinary medicine. Key concepts are reviewed, based on the broader experience in human medicine, with adaptations, as indicated, to veterinary medicine. That a paper has been published in a peer-reviewed journal does not guarantee its credibility; guidelines are given regarding the general merit of different kinds of articles, as well as checklists and criteria that can be used to assess a paper. Specific study designs, their merits and limitations, are briefly discussed. Standard numerical indices for assessment of studies involving treatments and for assessments of diagnostic tests are summarized. Criteria for assessing drug trials are presented. The principles of statistical analysis are described, including practical considerations and common errors. Finally, numerous sources of bias are reviewed.

Keywords: Evidence-based veterinary medicine; Experimental design; Clinical tests; Diagnostic tests; Statistical analysis

1. Introduction

In 1992, it was estimated that only 4% of therapeutic decisions in human medicine were based on strong evidence from clinical studies, 45% were based on minimal evidence from studies but strong clinical consensus, and the remaining 51% were based on personal opinion [1]. However, we are currently living in the 'information age'; new information is being discovered and communicated at an ever-increasing rate. Due to the current availability of information and the relative ease with which it can be accessed, leading-edge practitioners (in both medical and veterinary practice) have an unprecedented opportunity (indeed a responsibility) to incorporate current, accurate information into their day-to-day activities. The purpose of this paper is to briefly review key concepts regarding critical reading of the scientific literature (and other sources of information), to make informed decisions. Since this article is intended primarily for veterinary practitioners, it will emphasize evidence-based veterinary medicine, and will draw heavily on two sources that cover this topic from the perspectives of human [2] and veterinary [3] medicine, respectively. Where appropriate, the discussion has been changed to refer to animals (in lieu of humans) as patients.

2. Evidence-based medicine

The term 'evidence-based medicine' was coined by Sackett et al. [4]. The process of evidence-based medicine follows five key steps [4]:

* Tel.: +1 403 317 2236; fax: +1 403 382 3156.
E-mail address: Kastelicj@agr.gc.ca.

0093-691X/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.theriogenology.2006.04.017

(1) Identify a clinical problem and express it as an answerable question.
(2) Search for the best evidence to answer the question.
(3) Critically appraise the evidence for validity and clinical relevance.
(4) Integrate this appraisal with clinical experience to formulate the best decision for the clinical problem.
(5) Evaluate the practitioner's performance by relating clinical decisions to the best available evidence.

With the increasing prominence of evidence-based medicine, a similar approach is also being used in veterinary medicine. However, the primary difference between evidence-based medicine and evidence-based veterinary medicine is that in the latter, the emphasis must be necessarily placed on poorer sources of evidence [3].

A common misunderstanding is to equate evidence-based medicine with randomized clinical trials. However, less than 14% of published scientific articles are randomized trials, observational studies are overlooked and patient preferences, clinical circumstances and clinician's expertise are undervalued [5]. Thus, evidence-based medicine should rely on multiple sources of information.

To practice evidence-based medicine, the appropriate sequence of events is to ask the correct question, acquire the information, appraise its quality, apply the results, and ultimately act on the patient [3]. It is essential to start by asking the right question. Categorize the question being asked. Establish priorities, including what is the most important for the patient. Determine what question has the greatest benefit for the lowest cost (i.e. time and resources). When formulating a question, you should take into account the following [3,6]:

(1) The patient or the problem; the evidence should be as similar as possible to the current situation, taking into account age, breed, primary problem, and the population to which the patient belongs.
(2) The intervention or exposures must be defined to guide the choice of the appropriate study design; it could be a diagnosis, therapeutic intervention, prognostic factor, or exposure.
(3) The control group. Define the alternative; it may be one drug versus another drug, or one drug versus no treatment. It may be a comparison of two diagnostic tests. It is often useful to consider what you would do as an alternative (including doing nothing).
(4) The clinical outcome; it must be important enough to influence the clinical decision. This could involve the patient, the owner, or both. Define what you hope to accomplish, measure, improve or affect, and the timeframe during which you expect it to occur.

3. Assessing the validity and value of a publication

The peer-review system is far from perfect; unfortunately, many poor-quality papers are published in peer-reviewed journals. That a paper appears in a peer-reviewed journal is not a guarantee that it is credible and useful. In a recent article detailing errors and shortcomings in scientific papers, it was concluded that 51 of 67 (76%) of articles published in a well-recognized journal were flawed [7]. The following are common reasons why papers are rejected [2]: failure to examine an important scientific issue; lack of novelty; failure to test the stated hypothesis; inappropriate study design; compromised conduct of study (bias or confounding); inadequate sample size; no, inadequate or inappropriate controls; inappropriate statistical analysis; unjustified conclusions; conflict of interest; and poor writing.

It is noteworthy that not all reports are regarded as being of equal value. In general, articles are ranked in descending order of reliability as follows [8]: systematic reviews and meta-analyses, randomized clinical trials with definitive results, randomized clinical trials with non-definitive results, cohort studies, case–control studies, cross-sectional surveys, and case reports.

It has been stated that papers can be discounted even before you have read the results section [2]. As a reader, there are three preliminary questions that you should ask [2]:

(1) What was the impetus for the study and what hypothesis (if any) was tested? The introduction should include a brief explanation of what is known and how the authors propose to modify or extend current knowledge or to provide new information. There should be a clear objective (ideally a hypothesis), indicating what is being tested. It is noteworthy that some studies (e.g. qualitative research, case reports) are not expected to have a hypothesis.
(2) What was the study type? Primary studies include experiments, clinical trials and surveys, whereas secondary research includes reviews (systematic or non-systematic) and meta-analyses, clinical guidelines, decision analyses, and economic analyses.
(3) Was the design appropriate?

Once you have evaluated the paper according to the criteria noted above, and if it still holds your interest, you should investigate it further, using the following criteria [2]:

(1) Was the study original? Although replication is important to assure repeatability and to facilitate meta-analyses, what new approach or information was provided?
(2) What animals were used and were they similar or different to the animal(s) under your care? Are the age, breed, management, and clinical condition relevant? How were animals recruited, including the criteria used for inclusion and exclusion?
(3) Was the study design sensible? What specific intervention was under consideration and what was the control group? What outcome was measured and how was it measured?
(4) Was systematic bias avoided or minimized?
(5) Was the study 'blinded'?
(6) Were preliminary statistical questions addressed? Specifically, you should consider sample size, statistical power (was the difference 'clinically' significant), duration of the study, and completeness of follow-up.
(7) Based on the criteria listed above, you should be able to briefly summarize what was done; this will help to interpret the results and discussion and determine the reliability of the information.

4. Study designs

There are two different kinds of study designs, descriptive and explanatory [2,3]. For a descriptive study, observations are recorded, there is no control group, and you should not attempt to explain causation or derive conclusions regarding treatments. However, descriptive studies are useful to formulate hypotheses. In an explanatory study, it is typical to compare two or more groups, e.g. those with and without a specific disease, specific treatments or diagnostic tests. There are two main types of explanatory studies, experimental and observational. In experimental studies, the investigator determines the method of selection of animals and the interventions that they are to receive. For observational studies, groups are formed on the basis of the treatments given. The principal study designs have been described in detail [2,3] and are summarized in the following paragraphs.

4.1. Reviews and meta-analyses

Ideally, a specific question should be studied in several places by different teams of investigators; this forms the basis for a systematic review, a comprehensive survey of all primary studies of the highest level of evidence, systematically appraised and summarized according to explicit and reproducible methods. A meta-analysis combines several studies of a similar design and analyses them as if they were a single study. The optimal approach to meta-analysis is when individual investigators agree to pool their raw data (as if the pooled data had been concurrently collected) and recalculate the results [9]. The proper approach to a meta-analysis includes [10]:

(1) A carefully considered and detailed protocol is written before the start of the study.
(2) Eligibility criteria are defined a priori.
(3) Results are graphed on a common scale to visually assess heterogeneity.
(4) Appropriate statistical methods are used to combine data.
(5) A thorough sensitivity analysis is used to assess the robustness of combined estimates (using different assumptions and inclusion criteria).

Results of a meta-analysis are usually presented as a forest plot, with odds ratios and confidence intervals for the individual studies and for the combined data. An odds ratio = 1 indicates no effect, whereas odds ratios <1 and >1 indicate a decline and improvement, respectively. If all confidence intervals overlap, the results are compatible (homogeneity of the results) and it is likely justifiable to combine data. However, if the confidence intervals do not overlap (heterogeneity), this suggests significant differences among studies and the data should probably not be pooled. Heterogeneity may be due to differences between trials with respect to populations, methods or operator bias.

4.2. Randomized controlled trial

There are two kinds of randomized controlled trials, experimental laboratory studies and experimental clinical trials [3]. Laboratory studies use experimental animals in a controlled environment. The investigator has complete control over animal allocation and treatments. Although this provides the best evidence of cause or treatment effect, the results may lack real-world relevance due to the conditions under which the study was conducted. Experimental clinical trials typically use privately owned animals, in natural environments with naturally occurring disease. Although the researcher typically controls allocation of animals to groups, the animal's owner usually gives treatments. With good design and execution, an experimental clinical trial can give valuable evidence under field conditions [3].

A randomized controlled trial has two important features: at least two groups and patients randomly allocated to groups. The control group can receive a default treatment, a placebo or no treatment, whereas the treatment group receives the treatment or intervention under study. All groups are observed in an identical fashion for a specified interval and differences in outcome are attributed to the trial. Ideally, the study should be conducted as double-blind (neither the owner of the animal nor the investigator knows the assignment to group). This trial design is well suited to drug treatments, surgical procedures, and other interventions. The advantages of a randomized controlled trial are that it is generally the most powerful design, it should reduce the risk of bias (with valid randomization and appropriate blinding), it facilitates subsequent meta-analysis, and it provides the best assurance that the differences can be attributed to treatments. However, these trials are rare in veterinary medicine [3], they are typically expensive (may limit the numbers of animals used, the duration and extent of monitoring) and it may be unethical to withhold treatment.

Although a randomized controlled trial is generally regarded as the most robust design, it is certainly not the best design under certain circumstances. For example, it may not be practical if the number of animals needed to detect significant differences is prohibitively high. Furthermore, for some studies, other designs are more appropriate (e.g. cohort design to assess prognostic signs and cross-sectional survey to determine the validity of a diagnostic test or screening test).

In some cases, a study that is purported to be a randomized controlled trial does not meet the strict definition [2]; due to non-random allocation of animals to groups (because it was impossible, impractical or unethical), it should really be designated as an 'other controlled clinical trial' [2]. In some cases, allocation is less than completely random. For example, sequential allocation (first animal to one group, next animal to the other group), allocation by ear tag or identity (even numbers in the treatment group and odd numbers in the control group) are not considered valid as they allow the researcher to know which group an animal would be in before a definitive decision is made to allocate the animal to a group.

4.3. Cross-over design

In a cross-over design, animals are assigned to one of two treatments and followed over time to monitor the outcome of interest. Thereafter, they are switched to the other treatment. There may be a 'washout' period to minimize carryover effects of treatment. Although each animal acts as its own control, thereby reducing the number of animals needed and increasing the probability of detecting a difference, treatments with persistent actions may confound the results.

4.4. Observational studies

Observational studies are based on observations between groups, but the researcher does not control allocation to groups. Although this approach is less powerful than experimental studies, it may allow work to be done that would otherwise be too expensive to study experimentally. There are three basic kinds of observational study: cohort, cross-sectional, and case–control.

4.4.1. Cohort study

In a cohort study, animals exposed to a putative causal factor are followed over time and compared with another group not exposed to that factor. The two groups are monitored. Alternatively, two different treatments can be compared. It is important to carefully match the groups and to minimize differences between groups other than the factor of interest. This design is well suited to prognosis studies (predicting the outcome early in the course of the disease) [2] and for causation studies (to determine if a factor is related to development of a disease or condition). Cohort studies are more reliable than case–control studies, but cheaper than randomized controlled trials. They can be used to establish the timing and sequence of events and if done prospectively, data collection can be standardized. However, they can be difficult to conduct as blind treatments and to find cohorts that match for all variables, except that under study. They often take a long time to complete (resulting in higher attrition) and for rare diseases, recruitment of sufficient cases is often difficult.

4.4.2. Cross-sectional survey

In a cross-sectional survey, a representative sample of the whole population is sampled and two groups are identified (typically those with and without a specific disease, respectively). Both groups are similarly assessed and the data are used to determine relationships between exposure to a specific factor and the presence of the disease; this is usually expressed as an odds ratio (this is the only type of study that yields true prevalence rates). This design is ideal for evaluating a new diagnostic test, including screening tests that are intended to identify diseases at a presymptomatic stage [2]. Although these studies are usually inexpensive and easy to perform, it is generally not possible to assess temporal relationships and determine cause and effect.

4.4.3. Case–control study

In a case–control study, animals that have developed a disease are identified and their exposure to a suspected cause or risk factor is compared to that of control animals (without the disease); the results are expressed as an odds ratio (not possible to determine absolute risk). These studies may be used to evaluate interventions as well as associations and are useful for causation studies (to determine if a factor is related to development of a disease or condition). These studies are generally quick to perform, inexpensive, and are often the only practical method to study rare diseases or those with a long incubation period. However, it is difficult to match the control group and eliminate confounding variables and since data are collected retrospectively, there may be missing or poor-quality data.

4.5. Case reports and case series

Case reports and case series can provide valuable information and are commonly done in veterinary medicine [3]. A case report is a report on a single patient, whereas a case series is a collection of case reports on the clinical description of a specific condition or treatment of a condition. Although these are typically the least reliable form of evidence, they can provide valuable information about new and rare diseases.

5. Numerical indices

A study regarding treatment should have the following properties: clear objective, random allocation of animals to treatments, consistent treatment of groups, double-blind (both owners and clinicians unaware of assignment to group), most (typically $> 80\%$) animals accounted for at the end of the study, and adequate follow-up.

In order to assess the importance of the results of a treatment-based study, the following indices must be calculated [2,3]:

Relative risk reduction (RRR) = (CER − EER)/CER.
Absolute risk reduction (ARR) = CER − EER.
Number needed to treat (NNT) = 1/ARR.
Relative risk (RR) = EER/CER.

The CER is the control event rate (proportion of control animals demonstrating an effect) and EER is the experimental event rate (proportion of treated animals demonstrating an effect). The RRR is the proportion by which the treated group improves compared to the control group, whereas the ARR is the absolute difference between the control and experimental group. In most situations, the ARR is more useful than the RRR. The inverse of the ARR is the NNT, the number of patients that need to be treated to prevent one bad outcome. The RR is the ratio of the experimental event rate to the control event rate. An RR < 1 indicates that the event is less likely in the experimental group than in the control group, whereas an RR > 1 indicates that it is more likely in the experimental group. However, RR can be misleading when it deals with rare events; doubling the risk has little impact on the actual number of patients affected [9].

It is often difficult to establish cause and effect. The following questions are a good checklist for interpreting causality [2]. Is there experimental evidence, is it strong, consistent among studies, is there an appropriate temporal (cause preceded the outcome) and dose–response relationship, does the association make sense based on both epidemiology and biology, is the association specific, and is it analogous to a previously proven causal association? If there is no association (RR close to 1.0), there is no cause and effect. However, if there is an association, assessing the magnitude and precision of the relationship is the next step. If there is an important, precise RR, the Bradford–Hill Criteria for causation [11] should be used to determine causality. These criteria include: study design, strength of association, consistency, temporality, biologic plausibility, specificity, coherence, and the existence of analogies. Recommendations are categorized as: (A) good evidence for cause and effect; (B) fair evidence for cause and effect; (C) insufficient evidence to make a decision; (D) fair evidence against cause and effect; and (E) good evidence against cause and effect.

6. Assessment of a diagnostic test

When assessing a diagnostic test, both the sensitivity (probability of a positive test in an affected animal) and specificity (probability of a negative test in an unaffected animal) must be determined (and confidence intervals calculated). To determine sensitivity, a group of animals known to have the disease (a 'gold standard' reference is essential), ideally representative of all phases of the disease, must be sampled. For specificity, animals known to be free of the specific disease must be tested. Both groups of animals should be representative of the population for which the test is intended.

There are four possible outcomes of a diagnostic test: true positive (disease positive and test positive), false positive (disease negative, test positive), false negative (disease positive, test negative), and true negative (disease negative, test negative). If these four outcomes are designated as a, b, c, and d, respectively, the following indices can be calculated [2,3]:

Sensitivity = a/(a + c) [true positive rate; how good is the test at detecting animals with the disease].

Specificity = d/(b + d) [true negative rate; how good is the test at excluding animals without the disease].

Positive predictive value = a/(a + b) [post-test probability of a positive test; probability of the disease in an animal that tested positive].

Negative predictive value = d/(c + d) [post-test probability of a negative test; probability of no disease in an animal that tested negative].

Accuracy = (a + d)/(a + b + c + d) [probability that all tests were correct].

Pre-test probability = (a + c)/(a + b + c + d) [prevalence].

Pre-test odds = prevalence/(1 − prevalence).

Likelihood ratio of a positive test = sensitivity/(1 − specificity) [how much more likely is a positive test in an animal with versus without the disease].

Likelihood ratio of a negative test = 1 − sensitivity/specificity [how much more likely is a negative test in an animal without versus with the disease].

Post-test odds = pre-test odds × likelihood ratio.

Post-test probability = post-test odds/(1 + post-test odds).

The post-test odds and probability are used for interpreting the results of a diagnostic test (positive or negative), based on the characteristics of the test, as well as the prevalence of the disease.

7. Assessment of drug trials

Practitioners are frequently faced with information regarding a new drug; in many cases, the manufacturer will provide this information. The criteria used to assess a new drug have been previously discussed [2]. The pharmacokinetics and bioavailability should have been established by treatment of healthy animals, and ideally in animals with the disease. The strongest evidence of the value of the drug would be to have it studied in one or more randomized controlled trials; these trials should also detect common drug reactions. Rare (and often more serious reactions) are typically derived from reports of adverse drug reactions; ideally these are followed by case–control studies to detect associations. Peer-reviewed published articles are generally the most reliable source of information, whereas internal reports and 'data on file' should be read and scrutinized to determine their validity. The following points should be strongly considered [12]:

(1) What is the ultimate objective of treatment for this particular patient (cure, prevent recurrence, minimize complications, etc.)?
(2) Based on the best evidence available, what treatment (if any) is most appropriate?
(3) What is the treatment target (on what basis will discontinue or change treatment)?

Be cautious when 'surrogate' end points are used (defined as a relatively easily measured variable that predicts a rare or distant outcome of either a toxic stimulus or therapeutic outcome, but which of itself is not a direct measure of either harm or clinical benefit [2]). In pharmaceutical studies, common surrogate end points are serum concentrations of a drug or its metabolite, minimum inhibitory concentrations (MIC) of an antimicrobial agent against bacteria on an agar plate, gross appearance of tissues (e.g. gastric ulceration), and radiological appearance. Surrogate end points are attractive as they often substantially reduce sample size, duration of treatment and the cost of clinical trials and they can allow assessments where primary outcomes would be either too invasive or unethical. However, surrogate end points are often of little value to determine the real worth and validity of a treatment.

8. Statistical analyses

A general comprehension of statistical analysis is essential to critically assess a publication. Despite appropriate randomization, sometimes there are significant differences between groups. Have the authors verified that the groups were similar at the outset, and if not, have they adjusted for differences? Were there any statistical outliers, and if so, how were they handled? Were the data analyzed according to the original protocol, or were subgroupings and re-analysis employed retrospectively? What kind of data have been collected and were appropriate statistical analyses used? Be wary when a very unusual statistical test is used, when a more common test would have apparently been appropriate [2]. In general, parametric analysis is the preferred approach. The classical assumptions for parametric analysis include normal distribution, equal variance and independence. Did the authors verify that the distribution was normal? If the distribution was non-normal, it can usually be ignored if not extreme, transformed (with parametric analyses used if the transformed data are normal), or non-parametric analyses can be used if the assumptions cannot be achieved [13]. Were residual plots examined to confirm a random scatter? If the data are logically paired, was a paired test used? If there was a factorial design, was a factorial analysis used and if so, were the means and interaction reported? What kind of multiple range test was used to locate differences; was it highly conservative (e.g. Bonferroni) or highly liberal (e.g. Least Significant Difference)?

That data are independent is a key assumption of parametric analyses. In many cases, information is collected repeatedly from the same animals, creating an inherent lack of independence. Therefore, a conventional analysis of variance is inappropriate. At the least, a repeated-measures analysis of variance should be employed, with many journals now requiring a mixed-models analysis [14].

Proportional data (e.g. fertilization rates, pregnancy rates, cure rates) are often collected. These data can be analyzed by Chi square; however, this analysis should be used only for 'expected' values > 5; for data sets with small numbers of observations, a Fisher's exact test is more appropriate. In some cases, the experiment was done in replicates, and the rates (on a per-replicate basis) were analyzed with analysis of variance. However, since percentage data are not normally distributed, the standard approach is to convert the data to a proportion, do an arc sine transformation, and analyze the transformed data.

Correlation analyses are commonly used and frequently misused. A correlation analysis measures the strength to the linear association between two variables (but does not establish cause and effect). For a correlation analysis to be valid, the following assumptions should be met [2]: the data were normally distributed, structurally independent (not forced to vary with each other), and only a single pair of measurements were made on each subject. A correlation of less than 0.6 is seldom large enough to be of practical significance, despite its statistical significance [13]. Furthermore, two measures could be highly associated, but if there relationship was not linear, they will have low correlations.

The principle of regression analysis is to predict one end point (dependent variable) from one or more other end points (independent variables). Ideally, the data should cover a substantial range, with adequate coverage over the range. To choose potential independent variables, one approach is to determine the correlations between potential independent variables and the dependent variable, and to include in the multiple regression analysis, all independent variables that were correlated ($P < 0.15$) with the dependent variable [15]. However, once potential independent variables have been identified, correlations between all pairs of potential independent variables should be determined; if any of these correlations are significant, only the independent variable that is most highly correlated with the dependent variable should be used (to avoid intercorrelation among independent variables, that could invalidate the regression analysis [15]). Since a regression analysis is based on the best-fit relationship for the data set used to develop it, ideally, it should be subsequently tested on a novel data set, to determine its utility. However, this is rarely done.

A critical aspect of statistical analysis is the concept of the experimental unit, defined as the smallest unit in a study to which a treatment can be assigned and give a response independent of the responses of other experimental units [7]. Although an individual animal is often the experimental unit, if treatments are administered to animals in a pen (e.g. a specific treatment given in the feed), then the pen becomes the experimental unit, regardless of how many animals it contains. Similarly, in the case of in vitro fertilization, when two or more ova are incubated with sperm in a small droplet, then the droplet, and not the individual ova, constitutes the experimental unit.

Although the terms accuracy and precision may be incorrectly used interchangeably, they are not synonymous. Strictly defined, accuracy refers to correctness of an observed value or average, relative to the true value (i.e. the 'gold standard'). In contrast, precision refers to the repeatability or dispersion of the measurements. Precision is usually expressed as a 95% confidence interval; it is expected that the true value lies within this confidence interval 95% of the time (19 times out of 20).

The value of evidence is usually directly proportional to the power of the study, i.e. the ability to detect real differences. Power depends on the size of the study population, natural variation in the parameters studied, magnitude of the effect of the intervention, and the nature of the data (it is easier to detect differences with continuous versus categorical data, e.g. body weight versus pregnancy rate). Trials with small sample size are subject to a high β (Type II) error, i.e. probability of concluding that there is no difference when there is truly a difference. Most investigators accept a β error rate of 20%, corresponding to a study power (ability to detect real differences) of 80% (power = $1 - \beta$); β error rates > 20% are subject to high risks of false negatives [16]. The Type I error rate (α), otherwise known as the probability or '*P* value', is the probability of falsely concluding that two treatments are different, when they are actually not different. In most cases, $P < 0.05$ is considered significant; however, there may be valid reasons to use a *P* value that is either higher or lower [7]. Regardless, it is preferable if the actual *P* value is clearly stated, so that the readers can draw their own conclusions regarding the degree of statistical difference.

Trials with a binomial outcome (e.g. pregnant versus not pregnant, survived versus died), require very large numbers of animals to have reasonable statistical power. For a binomial distribution, the 95% confidence interval is calculated as follows [7]: [1.96] [square root ($p \times q/n$)], where p is the probability of one outcome, $q = 1 - p$, and n is the number of observations.

For a binomial distribution with a probability of 50%, the confidence intervals for 25, 100 and 200 observations are 30–70, 40–60 and 43–57%, respectively [7], clearly illustrating that large numbers are essential to reduce the width of the confidence intervals. As a practical example, if pregnancy rates for two stallions are calculated on the basis of <100 mares/stallion, unless the apparent difference in pregnancy rates is >15 percentage points, there is a >95% chance that the difference is not real [7]. Furthermore, if the difference is <10 percentage units, it will not be statistically significant unless the two values are each based on >190 units [7]. It is noteworthy that these confidence intervals do not take into account biological and other unaccounted for variation; these could easily contribute an additional 20% more variation [7].

9. Bias

Bias is the systematic variation of measurements from true values. There are several types of bias [9,10]:

(1) Selection bias: unequal assignment to treatments.
(2) Diagnostic (or detection) bias: owners of subjects avail their subjects to more examinations and tests.
(3) Recall bias: owners of affected patients are more likely to recall exposures or incidence than those of non-affected patients.
(4) Attrition of susceptibles or recency of market introduction: those that do well continue, whereas those that do not well discontinue. If clinicians believe a new product is safer, they may prescribe it to patients with increased risk of complications.
(5) Publication bias: significant results are more likely to be published.
(6) Language and citation bias: among published studies, those with significant results are more likely to get published in English, to be cited, and to be published repeatedly.
(7) Database bias: in less-developed countries, studies with significant results are more likely to get published in a journal indexed in a literature database.
(8) Inclusion bias: criteria for including studies in a meta-analysis may be influenced by knowledge of the results of the set of potential studies.

In addition to bias, there are numerous other conditions that jeopardize the conclusions of a study. Inadequate numbers of controls, inappropriate controls or non-contemporary controls (including historical controls, before versus after treatment, comparisons between different places, and comparisons between experiments or with other reports in the literature) greatly weaken a study.

10. Conclusions

This paper is a brief review of the key concepts of evidence-based veterinary medicine, derived largely from two recent monographs in this area [2,3], several publications in the primary literature, and the author's experiences as a scientific reviewer and journal editor. Although evidence-based veterinary medicine is a relatively new concept, it is increasing in prominence as a means of coping with a rapidly burgeoning body of scientific information and to increasing demands and expectations of clients that have unprecedented access to information (but may lack the context and expertise to make critical assessments). The purpose of this paper is to provide guidelines and tools to critically assess information, to improve clinical practice and client education.

References

[1] Field MJ, Lohr KN. Guidelines for clinical practice. Institute of Medicine, Washington, DC: National Academy Press; 1992. p. 34–9 [cited by Williams JK. Understanding evidence-based medicine: a primer. Am J Obstet Gynecol 2001;185:275–8].

[2] Greenhalgh T. How to read a paper. The basics of evidence based medicine. London, UK: BMJ Books, BMJ Publishing Group; 2001, 222 pp.

[3] Cockroft PD, Holmes MA. Handbook of evidence-based veterinary medicine. Oxford, UK: Blackwell Publishing Ltd.; 2003, 210 pp.

[4] Sackett DL, Rosenbert WMC, Muir Gray JA, Haynes RB, Richardson WS. Evidence based medicine; what it is and what it isn't. Br Med J 1996;312:71–2.

[5] Bhandari M, Giannoudis PV. Evidence-based medicine: what it is and what it is not. Injury 2006;37:302–6.

[6] Elphick HE, Smyth RL. Research: the principles of evidence-based medicine. Curr Pediatr 2004;14:525–31.

[7] Amann RP. Weaknesses in reports of "fertility" for horses and other species. Theriogenology 2005;63:698–715.

[8] Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature. IX. A method for grading health care recommendations. J Am Med Assoc 1995;274:1800–4.

[9] Williams JK. Understanding evidence-based medicine: a primer. Am J Obstet Gynecol 2001;185:275–8.

[10] Zou KH, Fielding JR, Ondategui-Parra S. What is evidence-based medicine? Acad Radiol 2004;11:127–33.

[11] Bradford-Hill A. Principles of medical statistics, 9th ed., New York, NY: Oxford University Press; 1971. p. 309–23 [cited by Williams JK. Understanding evidence-based medicine: a primer. Am J Obstet Gynecol 2001;185:275–8].

[12] Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology—a basic science for clinical medicine. London: Little Brown; 1991. p. 187–248 [cited by Greenhalgh T. How to read a paper. The basics of evidence based medicine. London, UK: BMJ Books, BMJ Publishing Group; 2001, 222 pp.].

[13] Healy MJR. What statistics does a paediatrician need to know. Curr Pediatr 2004;14:507–12.

[14] Littell RC, Henry PR, Ammerman CB. Statistical analysis of repeated measures data using SAS procedures. J Anim Sci 1998;76:1216–31.

[15] Cook RB, Couter GH, Kastelic JP. The testicular vascular cone, scrotal thermoregulation, and their relationship to sperm production and seminal quality in beef bulls. Theriogenology 1994;41:653–71.

[16] Lochner H, Bhanari M, Tornetta P. Type II error rates (beta errors) in randomized trials in orthopaedic trauma. J Bone Joint Surg Am 2001;83A:1650–5.