How to Set Up a Clinical Research Project: Analytical Science vs. Anecdotal Evidence

N. A. White, DVM, MS, Diplomate ACVS and A. J. Kane, DVM, MPVM, PhD

To discover the cause of a disease or to decide if a new treatment is effective requires using scientific method. Anything short of critical evaluation leaves the potential for recommending an ineffective treatment or applying an incorrect diagnosis and subsequently the wrong treatment. Proper clinical research takes time, and frequently appropriate financial support and technical resources are needed. Unfortunately there are no shortcuts to solving these problems, and our scrutiny in evaluating new ideas must be as vigorous as our search for them. Authors' addresses: Marion duPont Scott Equine Medical Center, VMRCVM-Virginia Tech, P.O. Box 1938, Leesburg, Virginia 20177-1938 (White); and Centers for Epidemiology and Animal Health, USDA:APHIS:VS, 555 South Howes St., Fort Collins, CO 80521 (Kane). © 2001 AAEP.

1. Descriptive Studies

The simplest form of research is recorded observations. The observer recognizes something different or unique from what has been previously reported. In veterinary medicine a particular disease or disease process is usually the focus of a case series, which describes a disease in detail and may draw limited conclusions about a treatment or the outcome. Normally to be published, the information needs to be new and add to the current knowledge on the subject (Table 1).

In this form of clinical research the primary focus does not test a hypothesis and usually these observations cannot establish superiority of one treatment over another unless there was no prior treatment. The cause may be suggested by simple observation, but usually Koch’s postulates are not fulfilled.

Descriptive studies can take one of the following forms:

1. Single observation—reported as a case report
2. Group of observations with similar characteristics or outcomes—reported as a series of cases.
3. Observations from a sample of a population—reported as a survey.

The data collected is normally the history, results of an examination, the treatment, and outcome. Any one or all of the data can be used for comparison to previous work or simply categorized or compiled to summarize the finding or observation. If numerical data is available, such as the results of a laboratory test, summary statistics—such as the mean, standard error, median, or rel-
An example of a descriptive study was entitled “Diagnostic findings and prognosis following arthroscopic treatment of subtle osteochondral lesions in the shoulder joint of horses: 15 Cases (1996–1999).”¹ This study resulted from the observation that radiographs of horses’ shoulders had subtle lesions that were later confirmed by arthroscopy to be cartilage and subchondral bone lesions in the glenoid of the scapula. In this study two variables were consistent for all the horses; a subtle radiographic lesion in the glenoid and arthroscopic evidence of a glenoid lesion. The authors felt that the radiographic results, arthroscopic results, and the outcome of the cases after surgery had not been reported before. Results of the lameness exam, intra-articular anesthesia, and scintigraphy, though helpful, were not as consistent as finding the radiographic lesions.

The retrospective (Table 2) study was completed by looking through radiographic and surgery logs to find all horses with subtle shoulder lesions and all horses that had shoulder arthroscopic surgery. The two sets of data were used to identify the cases with both subtle lesions and arthroscopic surgery on the cases. In all case series a common element must be identified and used for case selection or exclusion. The remaining material from the case records then became the variables to examine in these types of cases. Some examples of findings from the case records in this series of cases that had not been previously reported included lack of response to local anesthesia with shoulder lameness, increased isotope uptake in regions of the shoulder that were not affected, and return of 80% of horses to their original level of performance.

In another descriptive study on enteroliths (“Evaluation of enterolithasis in equids: 900 cases (1973–1996)”² a large number of cases was examined. The study derived valuable information about enteroliths and the animals involved; however, there is a limit to the statements that can be made about the cause of enteroliths or the population that was examined. Because this is a descriptive study, which examined only the horses with enteroliths, comparisons to another population cannot be made. The paper suggests that drinking water with high magnesium concentrations or eating alfalfa hay may be the reasons for the increased number of enteroliths seen in California. However, the research was not designed to answer that question. A control population was not studied to see if horses in California that do not have enteroliths also drink water with high magnesium concentrations or eat alfalfa hay. Descriptive studies such as the enterolith study cannot be used to prove a hypothesis without some type of comparison.

2. Experimental Studies
A hypothesis about a clinical problem or treatment can only be solved by experimentation using an analytic study rather than simple observation. The analytic design is the only way to answer questions about efficacy or if there is a true effect of some treatment or trait. The distinguishing characteristic of this type of analytical study is that the investigator interferes with the natural progression of events by controlling the environment and applying some treatment or intervention. The study is conducted so as to decrease the

---

Table 1. Steps for Completing a Descriptive Study

- State the question: What is it you want to prove or what is it you believe is different.
- Record specific findings from records.
- Clearly describe the criteria for case selection.
- Attempt to keep any variables the same when including cases.
- Describe as completely as possible the population from which cases are selected.
- Acknowledge the difficulty in running statistics unless the data is present for all cases.
- Use appropriate summary statistics to describe the data.
- Statistical comparisons or tests usually are not necessary to report these observations.

Table 2. Retrospective vs. Prospective Studies

Studies are best characterized by how the sample is selected. Whether the study is retrospective or prospective does nothing to indicate the type of study or what types of objectives the study can or cannot accomplish.

- If the study is designed after collecting the data, the study is a retrospective study.
- If the study is designed prior to collecting the data, the study is called a prospective study.
- The retrospective study may not be able to complete comparisons because no plan was made to ensure that all the data was collected for each case.
- Data for the prospective study is collected in real time so there is data present for every case, which is often not true for the retrospective study.
- When properly planned including deciding about any needed statistical analysis, the prospective study tends to produce more reliable data and is the preferred way to complete clinical research.
variables, which can affect the outcome. See Table 3 for a list of the steps involved in completing an experiment.

There are several important parts to the experiment that are needed to make the results valid. The hypothesis or expected result needs to be stated prior to the beginning of the study. A control or comparison group is needed to make sure there is a real effect in the treated group compared to the untreated group. If the evaluation of the outcome is subjective, such as the degree of lameness, the evaluator must not be aware of (or blinded to) which animal has been treated. If research involves a specific measure, such as quantitation of an antibiotic concentration, the technique must be calibrated against a standard. In some cases a measurement may be artificial because of the environment chosen to complete the experiment. An example of an artificial environment or measurement is oxygenation of horses under general anesthesia. In this case the oxygen tension may not represent what actually happens when a horse is breathing room air, but the measurement under these conditions may still be a valid test as long as the treated group and control group have measurements taken during anesthesia. Lastly, the number of samples needed or the number of animals needed to identify a significant difference needs to be determined. This can be determined by calculating the power for an experiment (see Table 4).

An example of a controlled experiment is in the article entitled “Effects of intravenous administration of sodium hyaluronate on carpal joints in exercising horses after arthroscopic surgery and osteochondral fragmentation.” In this study the question posed was whether sodium hyaluronate (HA) administered systemically could change the outcome in a model of osteochondral fragmentation. Six horses were in the experimental group with the same number in the control group. The environment was controlled so that the only variable or difference between the groups was the administration of the HA. To answer the question, known measures of joint inflammation were used during the study. The study found significant differences in prostaglandin E2 concentrations and protein concentrations in the joint fluid in horses treated with HA compared to controls. Because the quantitative measurements were objective, blinding was not necessary to draw a conclusion (see Table 5). However, blinding of the evaluator was necessary to arrive at unbiased lameness scores, which were also

Table 3. Steps for Completing an Experiment

- State hypothesis: What do you think is true that is different from what is current knowledge.
- What measurements will you make?
- Select experimental subjects, tests, cells, levels of a hormone, etc.
- Select controls to use for a comparison with the treatment group.
- Make sure there are enough experimental subjects included to identify a significant difference if one exists.
- Control the environment and change only one variable (the treatment under investigation) at a time.
- Allocation of treatment: systematic vs. haphazard vs. randomization
- Recording the results
  - Change from baseline to outcome, measuring the changing variable.
  - Compare treated or changed variable to the control animal or system.
  - Compare outcome measured between treatment group and control group.
- Perform statistical comparisons with the appropriate test: this may require a statistician to help design the experiment.
- Draw conclusions about the hypothesis: Did the test prove what you thought.

Table 4. Power of a Study and Sample Size

Statistical power is the ability of a study to demonstrate an effect if one truly exists. If a study fails to demonstrate a significant effect (e.g., no difference between treatment and control groups), then it is important to ask how much power the study had to detect an effect if one truly existed. Statistical power is primarily determined by four factors:

1. Sample size.
2. Variability of the measurements (random error).
3. Magnitude of the difference between groups being compared.
4. The desired level of statistical significance.

Studies with a very small number of subjects are said to have low power. They are unlikely to detect significant effects (differences between groups) unless the measurements are very precise and the magnitude of the effects are large. Studies with a huge number of subjects are said to have a lot of statistical power. They may detect statistically significant effects that are so small in magnitude that they are not clinically relevant. Prior to beginning a study it is a good idea to consult with an expert in statistics or study design to estimate the sample sizes needed to achieve reasonable power for that particular study. When critically evaluating a scientific work, remember to stop and consider how much power the study likely had. Look at the sample sizes and the magnitude of the difference between groups, not just statistical significance.
example of Viagra® which was tested in men with erectile dysfunction for at least 6 months. The subjects in the study received either Viagra or a placebo “sugar pill” and both the researchers and the men being treated were blind to which pill each subject was receiving. Viagra had a 25% placebo effect, or, 25% of the men receiving the placebo had a positive result. Because up to 82% of men taking the higher dose of Viagra had a positive effect, Viagra had a significantly greater effect than the placebo and was therefore determined to be an efficacious drug. The positive result found in the men taking the placebo is called the placebo effect.

This same placebo effect can occur in evaluating animal responses such as lameness. Owners often want a new treatment to work and will demonstrate a placebo effect by recording a positive result even when an ineffective treatment was administered. The placebo is used to detect bias by having the evaluator blinded to which subject received the treatment and which one received the placebo.

Another example of the importance of careful use of controls and blinding the evaluators is from research on the use of acupuncture in horses with navicular disease and laminitis (“Electroacupuncture in the treatment of chronic lameness in horses and ponies: a controlled clinical trial”). In this study the experimental animals were treated with acupuncture while the controls had all the procedures completed in the same way, including housing, but no acupuncture. The horses, those with chronic navicular disease or those with chronic laminitis, were randomized during the selection for being in the treated or control group. This was done by giving each horse a number and the selection completed using a random number generator. This was necessary to remove bias in the selection of horses that would receive the treatment. Evaluation for lameness during the study was done by a grading system and the evaluator was blinded to which horses were treated and which were controls. At the end of the study, equal numbers of treated and control horses improved in the lameness grade measured both by subjective grading and by force plate evaluation. Without the controls the project could have easily claimed that acupuncture was beneficial for navicular disease and chronic laminitis.

Statistical analysis is necessary in most research because measured differences are not often large enough to exclude the possibility that chance was responsible for the difference between treatment or control groups. In paired trials the statistical methods may be very simple, whereas in trials with repeated samples or more than one group or variable the statistics will be more difficult. Most researchers should consult an expert in statistics prior to beginning a study and again before the data are analyzed.
3. Observational Studies

Unlike an experiment where the researcher controls the environment and tries to influence the outcome with a treatment, observational studies are conducted by monitoring the normal course of events. They are usually conducted in a natural or clinical setting as opposed to an artificial laboratory environment, and the researcher does everything possible to avoid interfering with the events under study. Although useful in evaluating the efficacy of treatments, the randomized clinical trial is usually superior for establishing the benefits of one treatment over another. Observational studies are usually used to better understand how a specific factor under investigation (the exposure) causes disease (the outcome) or to identify where preventive measures may be helpful. Observational studies can also be used when it is difficult or unethical to artificially create a disease model in a controlled environment. Unlike an experiment such as a clinical trial, observational studies can be done prospectively or retrospectively. Therefore, they are often less expensive. In an observational study, exposures (or treatments) are not assigned to the study subjects and randomization cannot be used. Also, because they are not conducted in a controlled environment, more than one variable can be changing at a time. For these reasons, observational studies are more susceptible to sampling errors and bias than experimental studies and often require more sophisticated analyses to identify a real effect or cause.

Error

Errors found in analytic studies have specific names and can be identified and often controlled by proper study design (Table 6).

Random Error

Random error includes the errors inherent in selecting samples for a study as well as actually taking measurements. These include normal biological variation between subjects. For example, consider a researcher who wants to estimate the average toe angle of a population of horses at a given race-track. Every single horse on the backside could be measured, but that is not necessary. A more efficient approach would be to select a sample, only measure the horses selected, and use that average as an estimate for all the horses on the race-track. Because of the normal biological variation between horses, the average measurement from the sample would include a little variation that resulted from which horses the researcher happened to choose when taking the sample. Whenever an average measurement from a sample is used to estimate the true value for a population, this variation is called the sampling error. It is one component of the random error in the estimate. Another component of the random error is measurement error. In this example, a little variation probably occurs whenever a hoof gauge is used to measure hoof angle. Some hooves are rather uneven, so they don’t fit in the gauge very well, and each time the gauge is applied to the same hoof it might be applied slightly differently. If these differences create a little negative error one time and a little positive error the next time, they are considered measure-

<table>
<thead>
<tr>
<th>Table 6</th>
<th>Two Types of Error: Random Error and Systematic Error</th>
</tr>
</thead>
</table>

Typically in clinical research measurements of some clinical parameter are taken on a sample of subjects from the population. Whenever a measurement is taken or a sample is selected two types of error can occur: random error and/or systematic error (bias).

Random error is the portion of variation in a measurement that has no apparent connection to any other variable. It includes measurement errors and sampling errors.

- Commonly thought of as background noise in a measurement or errors due to chance.
- A measurement with little random error is said to be precise.
- The primary way to decrease random error is to increase sample size, the number of subjects being examined.

Systematic error, also called bias, is patterned variation of measurements away from the true value. Any consistent tendency away from the true value in measurement or sampling produces bias.

- Bias is sometimes used synonymously with subjective, however, seemingly objective sampling or measurement techniques can also be biased.
- Unbiased estimates are said to be valid estimates.
- Systematic error is not decreased by increasing sample size because the bias in the measurement is still present no matter how many subjects are evaluated.
- Systematic error can be minimized by improving study design or controlled during data analysis.
Systematic error, also called bias, is another type of error that can occur during sampling or during measurement. If the sample used to estimate toe angle was selected haphazardly, it could be biased. For example, consider the effect of an observer who, in selecting horses to measure, tended to avoid fractious looking horses that might not stand well while their toe angles were being measured. If these horses tended to be the younger horses on the racetrack and younger horses tended to have higher (or lower) hoof angles than older horses, the estimate from the sample would be biased by the horses’ ages. This bias was created by a systematic error in sampling. Measurements themselves can also be biased. If the scale on a particular hoof gauge was a little off in one direction, all the measurements being taken would be biased in the same direction. Now consider if this hoof gauge was used to measure horses in a treatment group and a different hoof gauge that measured a little off in the opposite direction was used to measure horses in a control group. It is easy to see how this study could be biased by the systematic errors in taking the measurements.

Confounding

Another type of error, which can bias the results of a study, is called confounding (see Table 7). Consider a hypothetical study of the effect transporting horses may have on the risk of colic. It may be that recent transport is the exposure being studied, and this has been identified as a risk factor that increases the number of horses with colic. Something else, which occurs at the same time or is associated with transport such as changing the diet prior to transport, is actually the factor that increases the risk of colic. In this example horse transportation was the exposure being investigated, but a change in diet prior to transport is a confounder that obscured the results of the study. These types of errors are difficult to avoid in the natural environment of an observational study. In order to determine if a treatment is truly effective or identify the real cause of a disease, confounding often needs to be controlled in the design or analysis of an observational study.

4. Two Types of Observational Studies

Case-Control Study

The simplest and most common type of analytic observational study to conduct in a clinical environment is the case-control study (Table 8). Like a case series the researcher starts with a group of cases that share similar characteristics or outcomes. The key difference is that the same data are collected on a group of control subjects that are then compared to the data from the actual cases. The controls should be as similar to the cases as possible in every way; except that they are not affected by the disease (outcome) under investigation. Selecting the subjects for the study (sampling) is based on the outcome, subjects, either meet the case definition and are included as cases, or it is known that they do not meet the case definition and are eligible as controls.

Appropriate control selection is a key part of designing a good case-control study. In some instances controls may need to be examined in exactly the same manner as cases to ensure that they are in fact free of the disease under study. For example,

---

**Table 7 Confounding**

Confounding, a special type of bias, is a distortion of the results, due to the influence of an extraneous factor on the relationship between the exposure and the outcome. Confounding is not due to errors in measurement or study design—it is due to a natural (usually biological) relationship between the exposure, the extraneous factor, and the outcome.

A confounder must meet three criteria:

1. Be associated with the exposure in the source population.
2. Be associated with the outcome.
3. Not be on the causal path between the exposure and the outcome.

Remember, just because an extraneous factor is associated with the outcome does not necessarily make it a confounder. It also has to be associated with the exposure being studied.

**Table 8 Key Elements of a Case-Control Study**

- Sampling is based on disease (outcome) status.
- First determine outcome status, then compare exposure status between two groups: cases and controls.
- Can be prospective or retrospective.
- Provide evidence to identify risk factors or support preventive recommendations.
- Includes cross-sectional, case-crossover, and proportional mortality studies.

**Advantages**

- Relatively inexpensive, quick, few animals needed.
- Good for screening large number of potential risk factors and a single outcome.
- Efficient for the study of rare diseases.

**Disadvantages**

- Cannot measure disease incidence because the entire population is not sampled.
- Improper selection of controls can easily introduce error.
- Sometimes difficult to establish temporal relationship between exposure and disease.
consider a hypothetical study of cardiac arrhythmias after endurance rides. It would probably be necessary to perform ECG exams on subjects selected as controls as well as the cases, just to make sure the controls do not in fact have any arrhythmias that are not readily apparent. In a study of colic, however, it would probably be fair to assume any horse under general observation after an endurance ride that did not show clinical signs of colic would be an appropriate control. It is important that the controls come from the same population that produced the cases. A good test is to ask this question about each control, “If this individual had in fact been affected by the outcome, would it have been included in the study as a case?” If the answer is yes, then most likely the individuals are in comparable groups. If the answer is no, then there is something inherently different about the cases and controls other than the outcome under study, and the groups are not comparable.

Regardless of how readily the disease/outcome status is determined, it is always essential that the controls are examined in the same manner as the cases to determine exposure status. In the study of arrhythmias for example, if electrolytes and blood gases were measured in the cases, then they would need to also be measured in the same manner in the controls. It would not be appropriate to assume the controls had normal values. Blinding is also appropriate at this point in a case-control study. If possible, researchers determining the exposure status should not be aware of whether the subject is a case or a control, particularly if it is a subjective determination.

The analysis of case-control studies can be relatively simple. Comparing the proportions of subjects exposed to a single risk factor between cases and controls can be done with a chi-square test. However, if the investigator wants to examine several risk factors at one time or adjust for a potential confounder, then multivariable methods such as logistic regression may be necessary. In either case it is a good idea for researchers to consult with an expert in statistics or study design before beginning the study. This avoids collecting all the data only to discover that they cannot be used to make valid comparisons. Table 8 lists the key elements of a case-control study.

An example of a published case-control study is “Horseshoe characteristics as possible risk factors for fatal musculoskeletal injury of Thoroughbred racehorses.”6 The idea for this study came out of a clinical setting from practitioners who felt that some types of traction devices on racehorse shoes increased the risk of catastrophic injury. Of all the horses dying on California racetracks, horses that died from catastrophic injury were selected for study (the cases). The controls came from the same population (horses dying on California racetracks) but only included horses dying from non-musculoskeletal causes of death. Each horse received a full necropsy so the outcome status was determined in the same manner for cases and controls. Also, because the hooves were removed before the shoes were examined, the exposure status (type of traction devices used) was determined in exactly the same manner for cases and controls; it was blinded. Dozens of different horseshoe characteristics were considered as potential risk factors, but only one outcome was analyzed at a time.

The case-control study design was a good choice for this study. Catastrophic injury is very rare. If horses had been selected for the study based on how they were shod (the exposure) the researchers would need to have examined the shoes of thousands of horses to collect the same number of cases (155) for study. By first selecting subjects based on the outcome (catastrophic injury), then selecting a similar comparison group, only 200 horses needed to be examined. The outcome status was determined by reviewing necropsy records from the past few years. The shoes, however, had been set aside and were not examined until after the study had been set up. So this study had both retrospective and prospective elements in the design. By comparing the data from horses affected by catastrophic injury with that from unaffected horses, horses shod with toe grabs were found to be at increased risk factor for catastrophic injury compared to horses with no toe grabs.

Cohort Study

Cohort studies can be conducted in a clinical environment, but they often take longer to complete and can be more complicated than case-control studies. In a cohort study, sampling starts with a defined cohort (group of subjects) from the population. The researcher first determines the exposure status, then follows this subset of the population through time and records when disease occurs. In the analysis, the occurrence of the outcome is compared between the two groups: exposed and unexposed. Like the case-control study, all the subjects should be examined in the same manner and the same data needs to be collected from all the subjects. It is a good idea to use blinding, so the person evaluating the outcome under investigation is unaware of the exposure status. If the exposed group was followed more closely than the unexposed group, a diagnostic bias would almost certainly develop. This means more cases could be diagnosed in the exposed group simply because of the increased scrutiny. Table 9 provides a list of the key elements of a cohort study.

Cohort designs are more efficient than case control studies when studying a limited number of exposures but several different outcomes. They are also more efficient if the outcome of interest is fairly common. If it is decided that several dozens of cases are needed for a meaningful comparison, the researcher is better off if there are only hundreds of subjects followed to see if they become cases rather than thousands or tens of thousands.
The analysis of cohort studies can be more complicated than experiments or case-control studies. Usually the incidence of disease is compared between exposed and unexposed groups. This can be done with simple statistical tests like the chi-square test, but often more complicated multivariable techniques such as logistic regression or survival analysis are used. It is best to consult with someone familiar with these methods before beginning this type of study.

The recently completed Yearling Radiograph Study is an example of a cohort design. In this study the presence of a radiographic lesion at the time of the yearling sales was the exposure. This was determined first using a collection of radiographs from the past few years. When the lesions were recorded the evaluator was unaware of the outcome status of the horse, and, therefore, was blinded. The outcomes (racing performance and the occurrence of clinical problems) were then measured using, respectively, historical records and a questionnaire mailed to owners. In the analysis, racing performance was compared between the group of horses with yearling lesions (the exposed cohort) and the group of horses without yearling lesions (the unexposed cohort).

Race records provided an objective, unbiased measure of racing performance. Because the questionnaire relied heavily on the owner’s recall of clinical problems, there was undoubtedly some bias present. Owners that knew they had purchased a horse with a radiographic lesion would probably be more likely to remember clinical problems that developed when this horse was racing. This type of bias is called recall bias. It is difficult to avoid when relying on memory of past events and should be accounted for in the interpretation of the results.

### References