

Review Article

Epidemiological study design and the advancement of equine health

G. T. FOSGATE* and N. D. COHEN†

Departments of Veterinary Integrative Biosciences, and †Large Animal Clinical Sciences, College of Veterinary Medicine and Biomedical Sciences, Texas A & M University, College Station, Texas, USA.

Keywords: horse; bias; case-control; causality; cohort; epidemiology; randomised control trials; study design

Summary

The primary purposes of epidemiological investigations are to learn about causal mechanisms related to disease incidence and identify factors for therapy and prevention. Epidemiological studies can be observational - further categorised as descriptive or analytical - or experimental. Investigators performing experimental studies, or randomised controlled trials (RCTs), randomly assign treatments or exposures to study participants for the expressed purpose of the study.

The most frequently encountered observational epidemiological studies employed to investigate issues of equine health are cohort, case-control, cross-sectional and case series. A cohort study is an investigation in which the researcher follows (observes) a group, termed the cohort, over time to measure incidence of a particular outcome. A case-control study is an investigation in which the researcher selects a group of affected individuals (cases) and a comparison group (controls) to investigate factors associated with being a case. A cross-sectional study investigates a group of individuals for study that is often defined by membership in a target population and data concerning outcome and exposure are either collected related to the same time-point or data are collected by investigators at the same time. Case series are descriptive studies used to generate hypotheses concerning predictors of disease or recovery that can be performed retrospectively or prospectively. The best evidence for clinical practice is often derived from patient-centred observational epidemiological studies and it is imperative that equine veterinarians become familiar with study designs for the appropriate interpretation of epidemiological findings.

Introduction

Epidemiology is a relatively young branch of science (Poole and Rothman 1998) that has been responsible for identifying important predictors of human (Feinstein 1988; Greenland *et al.* 2004) and equine (Archer and Proudman 2006; Parkin 2008) health. The

philosophy of science plays an important role in the design and conduct of epidemiological studies just as in other branches of science. Refutationist philosophy, attributed to the writings of Karl Popper (Popper 1963; Buck 1975; Weed 1985), is currently considered the most appropriate method of discovery in science.

Refutationist philosophy is based on using deductive logic to infer predictions from a hypothesis and then comparing observations from the study to those predictions. It encourages scientists to consider multiple hypotheses and collect evidence to refute them one at a time. Every genuine test of a scientific hypothesis should be a test to refute it rather than to support it. A hypothesis may be refuted by a single study in which the collected data are not consistent with the hypothesis and contradict predictions (if collected data are not erroneous). Failure to refute a hypothesis is considered evidence in favour of that hypothesis; however, it is considered impossible to prove a hypothesis definitively. Scientists are therefore left with what can be considered the best guess given the scientific evidence collected to date.

According to the coherence theory of truth (Young 2001), this 'best guess' could be considered scientific truth, but it might be better to consider the findings of scientific endeavours as useful, evolving ideas that allow us to better understand our world. A simple example of this notion is the increasing knowledge concerning the shape of the Earth. Originally the Earth was considered flat based on the methods of observation available and this was a reasonable conclusion and an acceptable approximation at that time. The Earth was later determined to be spherical and improved measurement instruments have since determined that it is an oblate spheroid (rounded with bulging around the equator) that changes shape over time (Dickey *et al.* 2002). Another example is refutation of the theory of spontaneous generation and the subsequent development of the germ theory (King 1983), which has been responsible for many advances in science and health.

Epidemiology is employed to better understand the distribution and determinants of disease and health states in animals and people (Kelsey 1996, p 3; Last *et al.* 2001). A feature that distinguishes epidemiology from other health research is the fact that observations are made directly on the population of interest. Observations on free-living populations often do not

*Author to whom correspondence should be addressed.

[Paper received for publication 24.06.08; Accepted 13.08.08]

allow for control of factors external to the study hypothesis as is possible in laboratory studies using disease models. The inability to control for all factors and directly observe causal mechanisms is the root of a controversy concerning whether risk factor epidemiology should be considered a science (Skrabaneck 1994; Weed 1998; Miettinen 1999; Greenland *et al.* 2004). The science of epidemiology, however, lies within the theories and methods used to obtain valid findings in the absence of having complete control over the exposures of the study population(s). Disagreement and confusion related to study design classification are not uncommon (Kramer and Boivin 1987; Grimes and Schulz 2002d; Marshall 2004), despite the fact that the ability to implement valid studies starts with a firm understanding of epidemiological study design.

Planning and implementing valid epidemiological studies includes consideration of the philosophy of science and specifically the refutationist approach to discovery. The framework for all investigations is the epidemiological study design, which is the blueprint from which the study will be performed. Research hypotheses should be formulated and study objectives specified explicitly prior to design of the study. The blueprint should not be modified either during data collection or after completing statistical analysis. A hypothesis stated after recognition of statistical significance does not contain the same credence as a hypothesis preceding data collection and analysis according to the refutationist philosophy (because predictions did not precede data collection). Therefore, studies can be considered as hypothesis-generating or hypothesis-testing, depending upon whether or not the study was designed to specifically test a particular research hypothesis. The quality of evidence is not the same for hypothesis-generating studies as it is for hypothesis-generating studies, irrespective of P values and other statistical findings: studies designed to test specific hypotheses provide higher quality evidence compared to hypothesis-generating studies. A scientist performing epidemiological studies to test specific hypotheses should have a primary focus on collecting valid data and making appropriate conclusions.

With the burgeoning interest in, and commitment to, the practice of evidence-based equine medicine, epidemiological study design is no longer exclusively the purview of scientists conducting these studies. It is generally regarded that patient-centred research for the species of interest provides stronger clinical evidence than do experimental studies, particularly experimental results extrapolated from heterologous species. Well-designed randomised, controlled trials (RCTs) are generally

considered the strongest form of clinical evidence (Feinstein 1988; McKee *et al.* 1999; Stephenson and Babiker 2000; Gellerstedt 2002; Grimes and Schulz 2002d) but observational studies can sometimes provide a similar level of quality (Benson and Hartz 2000; Concato *et al.* 2000). RCTs are sparsely reported in equine medicine and observational epidemiological studies represent the most common and important source of clinical evidence for equine practitioners. Because of these circumstances, equine practitioners must also be familiar with the principles of epidemiological study design.

The purposes of this report are to review the basic designs of epidemiological studies and to provide guiding principles for interpretation of results from these studies. Some recent equine epidemiological studies are cited to exemplify certain strengths and limitations of which readers should be aware. Our intent is not to denigrate anyone's work: limitations are inherent to the study of free-living populations, and both contributors of this report have authored papers with ample examples of such limitations. Included articles were chosen because they were recently published in this journal and they are good examples of the strengths and weaknesses of various study designs. A glossary of epidemiological terms is provided in the appendix to facilitate the reading of this report.

Epidemiological study designs

Overview

The purpose of an epidemiological study is to estimate a valid measure of association (e.g. an odds ratio) or other population parameter that will be precise upon statistical evaluation (Fig 1). An estimate of a parameter is considered valid if the expected value (over infinite replications) is the true value (Brenner 1991; Mertens 1993; Szklo and Nieto 2000, pp 125-126). Bias is the lack of validity, and a study is considered biased if a systematic error is present in the study design, data collection or data analysis (Brenner 1991; Flegal *et al.* 1991; Rothman 2002, pp 94-95). A systematic error is a persistent error having a nonzero mean that cannot be attributed to imprecision in the system of measurement (Fosgate 2006). Random error is due to imprecision in a measuring instrument or protocol used to collect data. A random error in the absence of systematic error will not typically cause bias. The effect of random, but not systematic, errors can be reduced by increasing the sample size. The true validity of a study, however, cannot be known because infinite replications are

TABLE 1: Basic features of the 3 general types of systematic errors that affect epidemiological studies

	Selection	Information	Confounding
Definition	Study population does not represent the target population in respect to important characteristics.	Information does not represent the true state of nature*.	Measured association does not represent the true association because of the effect of another factor.
Source	Error in study subject participation.	Error in collection of information from study subjects.	Error due to failing to account for the effect of another factor.
Effect	Results might be internally valid but do not represent the population of interest.	Results could be biased towards or away from the null (no effect) value.	Results could be biased towards or away from the null (no effect) value.
Prevention	Employ appropriate selection techniques for study subjects.	Standardised collection of objective data and blinding of data collectors.	Matching and restriction procedures.
Control	Typically not possible to control once study has been performed without subsequent validation procedures.	Typically not possible to control once study has been performed without subsequent validation procedures.	Effects can be reduced using statistical methods if appropriate data are collected and available for analysis.

*Information bias that is differential between comparison groups is a study bias but nondifferential errors could be considered random classification errors.

never performed and the true population value is typically unknown. Therefore, investigators must use logic to reason if epidemiological studies contain systematic errors that could cause findings to deviate from the true state of nature. The study design is one of the most important considerations when reasoning about the presence or absence of systematic errors because some designs are more likely to contain certain errors that can cause substantial bias.

The 3 general categories of systematic errors that can affect epidemiological studies are selection, information and confounding bias (Ibrahim and Spitzer 1979); further discussions can be found elsewhere (Sackett 1979; Miettinen and Cook 1981; Greenland and Morgenstern 2001; Grimes and Schulz 2002a; Christenfeld *et al.* 2004) (Table 1). Misclassification (Copeland

et al. 1977; Gladen and Rogan 1979; Birkett 1992; Brenner and Loomis 1994; Jurek *et al.* 2008) is a common form of information bias that can affect classification of either exposure or outcome. Nondifferential misclassification of exposure information is an error that is independent of outcome status: the direction and magnitude is equal for those with and without the outcome. Nondifferential misclassification of the outcome is an error that is equal across exposure categories. Differential misclassification of the outcome or exposure is an error whose magnitude is different within the different categories of the other variable (e.g. errors associated with exposure classification are different within diseased and nondiseased groups). Statistical tests used to make inferences about hypotheses are based on random error within data that do not incorporate measures of study validity.

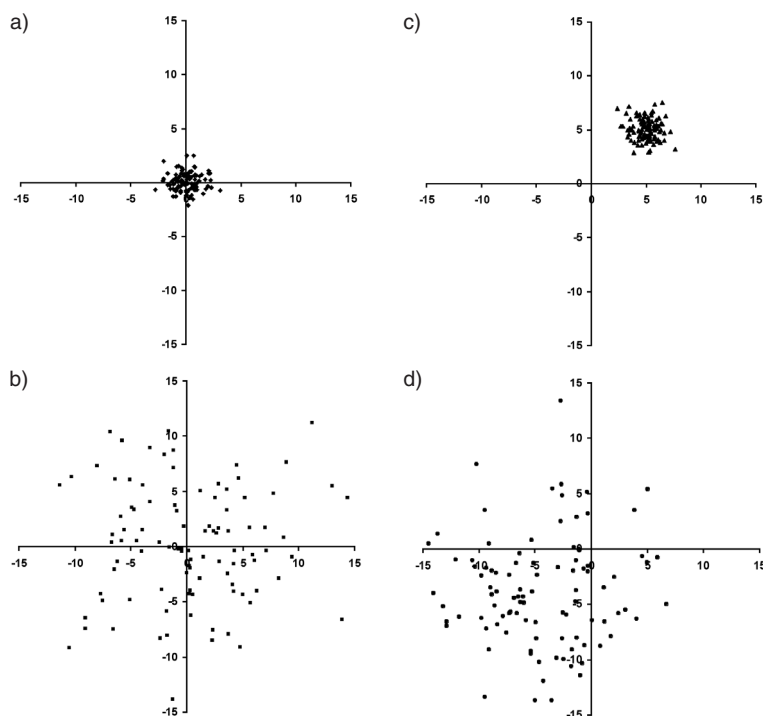


Fig 1: The effect of bias and precision on epidemiological measurements. The origin (0,0) is considered to be the true value. Values simulated from a precise and valid (unbiased) system (a), an imprecise and valid system (b), a precise and invalid (biased) system (c), and an imprecise and biased system (d).

TABLE 2: General classification scheme for frequently-encountered epidemiological studies

Design	Classification criteria			
	1: Data level	2: Exposure assignment	3: Sampling of subjects	4: Follow-up
Randomised clinical trial (RCT)	Collected at the individual level.	Randomly assigned by investigators.	Based on characteristics that make them representative.	Data collected for multiple time-points.
Prospective cohort study	Collected at the individual level.	Determined by factors other than by random assignment.	Based on a characteristic in the present.	Data collected for multiple time-points.
Retrospective cohort study	Collected at the individual level.	Determined by factors other than by random assignment.	Based on a characteristic from the past.	Data collected for multiple time-points.
Cross-sectional study	Collected at the individual level.	Determined by factors other than by random assignment.	Based on membership in a specific group.	Data collected at or related to a single time-point.
Case-control study	Collected at the individual level.	Determined by factors other than by random assignment.	Based on prior knowledge of outcome status (case or control).	Data could be collected related to a single or multiple time-points.
Case series	Collected at the individual level.	No predetermined exposure of interest.	Based on characteristics that make them representative.	Data typically related to multiple time-points.
Ecological study	Collected at the aggregate level.	Determined by factors other than by random assignment.	Based on membership in a specific group or population.	Data typically related to a single period of time.

Epidemiological studies can be observational or experimental; the former can be further categorised as descriptive or analytic. Reviews of basic designs considering strengths and weaknesses have been published related to human conditions (Chuang and Reizner 1993; Balkau and Eschwege 1995; Freudenheim 1999; Harris 2000; Stephenson and Babiker 2000; Baxter 2001; Grimes and Schulz 2002a). Observational studies are ones in which the investigator does not assign a treatment or exposure to the study participants for the expressed purpose of the research. The treatment or exposure status of the individual is determined by factors unrelated to the study objective and subjects are simply observed for development of the outcome (e.g. disease) based on self-selected (or otherwise determined) group membership. Investigators performing experimental studies, or randomised controlled trials (RCTs), randomly assign treatments or exposures to study participants for the expressed purpose of the study.

Descriptive studies are considered hypothesis-generating and are implemented without *a priori* specification of a hypothesis for testing. Analytical studies are designed to test specific research hypotheses. The use of statistical hypothesis testing *post hoc* (i.e. after the collection of data) does not imply that that study was analytical when designed and implemented.

The base epidemiological study design (framework) is defined by 4 factors (Table 2; Fig 2): 1) the level at which data are collected (individual vs. aggregate); 2) the method(s) of creating exposure groups for comparison; 3) the sampling process to identify individuals for study, and, 4) the follow-up procedures for data collection and outcome determination. The term individual could refer to individual animals or aggregates of animals if the unit of

interest is actually the aggregate (e.g. farm). The most frequently encountered observational epidemiological studies employed to investigate issues of equine health are cohort, case-control, cross-sectional, and case series (Table 3). Ecological studies are based on measurements made at the group (aggregate) level even though inferences are desired to be made for the individual. Instances when ecological studies would be appropriate in equine epidemiology are exceedingly rare. Experimental studies must be analytical and are typically clinical trials of patients designed to investigate therapeutic or preventive interventions.

The quality of evidence towards determining causality that a study provides depends upon the selected design and its implementation. Investigators must provide adequate details when authoring manuscripts that allow for subsequent evaluation and compilation with other study results to evaluate cumulative evidence. Standards have been developed for the reporting of studies related to human health to facilitate assessment of quality and future aggregation of evidence. These guidelines are also useful during the design of epidemiological studies to ensure good data quality after completion of the research. Standards have been published related to RCTs in general (Altman *et al.* 2001; Moher *et al.* 2001), equivalence or noninferiority RCTs (Piaggio *et al.* 2006), an extension to nonpharmacological treatments (Boutron *et al.* 2008), and specifically for reporting negative consequences of these studies (Ioannidis *et al.* 2004). Similar guidelines have also been produced for the reporting of results from diagnostic test evaluations (Bossuyt *et al.* 2003), nonrandomised intervention evaluations (Des Jarlais *et al.* 2004), and observational epidemiological studies (Vandenbroucke *et al.* 2007).

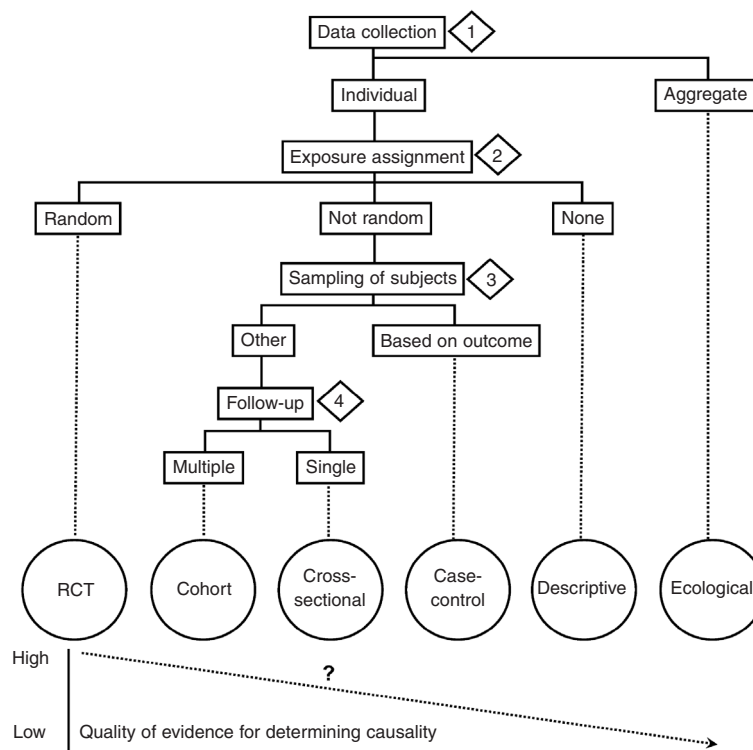


Fig 2: Flow chart for classification of epidemiological study designs. Study design is determined by: 1) whether data are collected at the individual or aggregate level; 2) how the primary exposure/treatment is assigned to study participants; 3) whether or not information related to outcome is used to select subjects for participation; and 4) whether data were collected related to a single or multiple time-points. Evidence for determining causality generally decreases as moving from left to right. When a cross-sectional study cannot determine that exposure preceded development of the outcome then its quality of evidence for determining causality should be ranked lower than case-control studies.

Experimental

General description: An experimental study is an investigation in which the researcher randomly assigns study participants to 2 or more treatment or exposure groups to test the hypothesis that the groups will be different with respect to the incidence of a particular outcome (e.g. disease or recovery from disease). If assignment of the exposure/treatment is not random then the study typically should be considered observational rather than experimental. The study design is defined by the factors: 1) data are collected at the individual level; 2) animals are randomly assigned into intervention/exposure groups; 3) animals are sampled to represent the target population of interest; and 4) outcome information is collected prospectively for a minimum of 2 times (once to confirm freedom of the outcome at the beginning and once at the end of the study).

Experimental studies (Rothwell 2005) (i.e. RCTs) are frequently expensive and difficult to design well but on average will provide the highest quality of evidence for identifying true causal associations (Horwitz 1987; Feinstein 1988; Rubinfeld 1995). Randomisation is performed in an effort to make groups comparable with respect to known and unknown factors that are independent predictors of the outcome. Confounding bias can occur when independent predictors of the outcome have a differential distribution across the comparison groups. Random assignment also prevents allocation bias (a type of confounding), which occurs when characteristics of study participants that predict the outcome (e.g. more severe cases) also affect treatment group assignment. The advantages of experimental studies are that the possibility of bias is reduced, incidence can be measured, and multiple endpoints or outcomes can be investigated. The disadvantages of these studies are that they can only be performed for ethical interventions with willing subjects, they can be technically difficult to design and implement appropriately, they only allow the investigation of a single intervention, and they can be cost-prohibitive.

A well-designed and implemented experimental study is less prone to systematic errors than other study designs and this

explains why results are considered the best epidemiological evidence for determining causal associations and for assessing clinical performance of a treatment or procedure. Selection bias, in general, occurs when the subjects selected for study are systematically different from the target population in respect to their response to the intervention or exposure. This is not a common source of error in experimental studies but can occur when enrolled participants are systematically different with respect to the outcome or their response to the intervention. Experimental studies reduce the extent of this error by incorporating criteria for inclusion and exclusion of participants. As a general rule selection bias in a study employing random assignment to treatment groups would limit external validity (generalisability) but results would still be internally valid for the enrolled subjects.

Information bias is a concern in all epidemiological studies. This error develops when collected information does not represent true information and is a particular concern when the error in data collection is different based on intervention group. Objective data collected using standardised (preferably validated) methods can reduce these errors. Data collection should also be performed in a blinded (masked) manner to prevent differential errors in collection from occurring. Selection and information bias should be minimised using appropriate study design procedures because *post hoc* adjustments require additional assumptions and are frequently difficult if not impossible to perform.

All epidemiological studies should be considered to contain some degree of confounding that can lead to bias if it is not appropriately controlled during the statistical analysis. Random assignment in experimental studies minimises the possibility of confounding, but randomisation does not ensure that treatment groups will be comparable for all characteristics, and differences among groups may occur following randomisation, especially when the sample size is small.

Study example: A recent example of an experimental epidemiological study is the paper by Coomer *et al.* (2007) entitled "Do subcutaneous sutures increase the risk of laparotomy wound

TABLE 3: Basic features of the most common observational study designs employed to investigate issues of equine health

	Case series	Case-control	Cross-sectional	Cohort
Study population	Diagnosed cases at a specified location during a specified time.	Cases selected from a specified location and controls selected to represent source population of cases (typically undefined).	Typically a well-defined population in respect to geography and time.	Well-defined group selected based on a shared characteristic.
Outcome determination	Animals have outcome as a requirement of entry into the study.	Outcome status known prior to entry into the study.	Outcome determined at single point in time.	Outcome determined over multiple time-points.
Exposure determination	No exposures	Exposure determined by investigators after outcome has occurred.	Exposure determined at the same time as the outcome.	Exposure determined prior to development of outcome.
Descriptive/analytical	Descriptive only	Analytical (hypotheses could be vague).	Descriptive or analytical	Typically analytical but could also be descriptive.
Measure of association	None	Odds ratio	Prevalence ratio	Risk ratio, rate ratio
Benefits	Relatively easy and inexpensive to perform.	Relatively easy and inexpensive to perform, results can be available in a short period of time if retrospective cases, efficient design for rare outcomes.	Relatively easy and inexpensive to perform, results available in a short period of time.	Can measure incidence of outcome and determine that exposure preceded outcome, relatively strong evidence towards causality.
Weaknesses	Not valuable for testing hypotheses related to causation.	Study design is susceptible to many biases and therefore evidence for causality might be weak.	Not possible to know if exposure preceded outcome and evidence for causality might be weak (except for fixed factors, e.g. sex).	Prospective follow-up is typically expensive and time consuming, not efficient for rare outcomes.

suppuration?" The objectives were "To determine if abandoning the use of subcutaneous sutures in laparotomy wound closure is safe and whether it reduces the risk of suppuration." Inherent in the authors' statement of objective is the comparison of the proportion (risk) of suppuration in laparotomy cases during the post operative period. The scientific hypothesis that was tested was that the risk of suppuration is lower in horses that do not receive subcutaneous sutures. The statistical null hypothesis was that there is no difference in the risk of suppuration between the 2 groups (subcutaneous sutures yes/no). The data collected in this study were not statistically inconsistent with the null hypothesis and therefore did not provide strong support for the research hypothesis.

The study employed a relatively large sample size and was adequately powered to detect an odds ratio of 2.4 or greater as statistically significant. The authors found 3-layer closures to have a 1.4 times higher odds of suppuration compared to 2-layer closures. The corresponding risk ratio based on the presented data is 1.3. The odds ratio when it is not equal to 1 will always be greater than the corresponding risk ratio (i.e. further away from the null value of 1), thus - if appropriate - the risk ratio should be presented. In this particular situation because the incidence of suppuration was low there was not much difference between the 2 measures.

Statistical P values and confidence intervals are measures of precision but validity must be qualitatively evaluated through assessments of the study design. Readers evaluating results of RCTs should examine the reports closely for the 3 general categories of bias, *viz.*, selection, information, and confounding. Selection bias can develop when included cases differ systematically from the population to which the results are expected to represent. The inclusion and exclusion criteria employed in this study did not suggest that this would be a concern. Examples of selection bias would include considering only horses having elective abdominal surgeries (which would exclude emergency surgery for colic), or if the study population excluded horses with strangulating colic lesions (when these cases might be expected to have greater and perhaps different risk factors for post operative infection). Selection bias in this situation would act to reduce the external validity of the study results.

Information bias is likely to occur to some degree in all studies and is particularly damaging to validity when the degree of error is different across comparison groups (e.g. different for 2- vs. 3-layer groups). Blinding or masking can be employed in experimental studies to prevent differential misclassification of the outcome. The extent that blinding was accomplished was not discussed in the report, and represents an important consideration because suppuration could have a subjective determination especially when outcome information was obtained from owners via telephone interviews.

Confounding is not likely to have resulted in substantial bias because of random assignment and the relatively large size of the study. However, it would have been valuable to have baseline data presented concerning potential predictors of suppuration including age, diagnosis, lesion type/anatomic location, procedure performed (resection and anastomosis vs. not), and duration of surgery.

Cohort

General description: A cohort study (Grimes and Schulz 2002c) is an investigation in which the researcher follows (observes) a group, termed the cohort, over time to measure incidence of a particular outcome. Selection of the cohort for study is performed

based on a characteristic shared among the individuals. This characteristic may be an exposure (or lack of an exposure) of interest but it can be any characteristic that allows enumeration of the cohort. These studies can also start with 2 distinct cohorts, the first exposed and the other unexposed, for the comparison of incidence between them.

Cohort studies can be descriptive or analytical, and can be defined as prospective, retrospective or a combination of both. A prospective cohort is defined based on a characteristic in the present and the observations take place in the present tense (i.e. from the present prospectively into the future). A retrospective cohort is defined by a past characteristic and the observation of the cohort is performed using pre-existing records to determine exposure status and incidence of the outcome through time (i.e. from a past time-point forward to some other time). Some studies have both prospective and retrospective cohort elements. In these studies the cohort is determined by a past characteristic and is to be followed from that time-point into the future. A cohort study is defined by the following features: 1) data are collected at the individual level; 2) exposure groups for comparison are defined by factors independent of the research; 3) animals are sampled based on membership in a cohort that is defined by a characteristic of the past or present; and 4) data concerning outcome status must be available for a minimum of 2 time-points (beginning and end of follow-up).

The advantages of cohort studies are that the incidence of the outcome can be measured, true risks can be calculated, the purported cause can be shown to have preceded the measured outcome, and the relationship between multiple exposures and multiple outcomes can be investigated (dependent upon the definition of the cohort and the data collected). Disadvantages of cohort studies include that they are relatively expensive to perform, rare outcomes might not develop at a high frequency, extended follow-up might be necessary, and they might have high level of censorship, or loss to follow-up. An advantage of retrospective cohort studies is that the follow-up period occurred prior to study initiation so large groups can be followed over long time-periods with minimal time investment by researchers. Retrospective cohort studies have disadvantages and it is important that accurate records exist to provide quality data. Additional limitations of this study design include that the data are not specifically collected for the purpose of the research (otherwise it would have been a prospective design), information about important variables might not be available, a large proportion of the cohort could have missing information, and large numbers of individuals can be lost during follow-up.

Cohort studies are generally considered more efficient for relatively frequent outcomes because rare events require large sample sizes and long study durations to obtain the necessary number of outcomes required for statistically significant differences. Cohort studies are not very common in equine epidemiology although many clinical studies could be defined as descriptive cohort studies.

Prospective study example: A recent example of a prospective cohort study is one by Pinchbeck *et al.* (2004) entitled "A prospective cohort study to investigate risk factors for horse falls in UK hurdle and steeplechase racing." The objectives of this study were "To identify and quantify risk factors for horse falls in National Hunt (NH) racing and to report the frequency of falling and falling-associated fatalities." Inherent in the authors' statement

of objective is the comparison of the proportion (risk) of falls between different levels of each potential risk factor. The scientific hypothesis tested was that the risk of falls is different among levels of at least some variables. The statistical null hypothesis was that there is no difference in the risk between the levels. Identified risk factors adjusted for race-track using a random effect included race-type (steeplechase vs. hurdle), going, weather, novice race, locomotion in the parade ring (walking calmly vs. other), and journey time to race. The study employed a relatively large sample size ($n = 2879$ starts among 2216 unique starters) but the relatively rare outcome of falls (124; 4.3%) might have prevented the authors from determining other significant predictors.

All epidemiological studies are expected to contain some degree of bias. The study population included horses racing at 6 tracks in the UK that were willing to participate. These tracks might not be representative of all race-tracks, but there was no reason to believe that selection bias would be a major concern. Information bias is likely a problem to some extent in all studies but for this investigation all data were either collected prior to the start of the race or from objective sources and therefore errors would probably not have varied based on outcome (fall vs. no fall) and would have been expected to be nondifferential. Confounding by unknown and unmeasured factors is always a concern but there is no indication that the authors excluded known predictors of falls that might have caused substantial bias in presented results.

Retrospective study example: An example of a retrospective cohort study is the paper by Lam *et al.* (2007) entitled "Descriptive analysis of retirement of Thoroughbred racehorses due to tendon injuries at the Hong Kong Jockey Club (1992–2004)." The objectives of the study were "To describe the frequency and pattern of retirements associated with SDF tendon injuries in Thoroughbred racehorses and to compare the characteristics of these horses with those that retired for other reasons." Inherent in the authors' statement of objective is the anticipation that horses that retire due to superficial digital flexor (SDF) tendon injuries will have some characteristics different from horses retiring for other reasons.

The study followed a large cohort from which 3727 horses were identified as having retired from racing with 510 (14%) retired due to injuries of the SDF tendon. Horses that were retired due to SDF injuries were found to be younger, spent less time in training, had fewer race starts and earned less money. The authors did not perform statistical testing because they sampled the entire population and were not interested in making generalised statements related to other racehorses. While this is a valid argument, the reporting of confidence intervals and P values allows interested readers to judge the level of precision of the data. An argument could also be made that scientific advances only occur when findings are useful for making generalising statements. The authors themselves appear to generalise some of their findings in the discussion section; an example is the suggestion that modification to early training regimens could reduce the number of horses retiring due to SDF tendon injuries.

In addition to the quantification of random error, all studies should discuss the possibility of systematic error. Selection and information bias should not be a concern for the aforementioned study, unless horses that were excluded from the database or that had missing data were systematically different from the horses with complete data included in the study. No information was presented concerning data quality or other validation procedures,

so it is not possible to critically evaluate this aspect of the study. The authors did not perform any statistical modelling to investigate and control for confounding. A particular concern would be confounding by age of the relationship between SDF injuries and career earnings. This is a concern because some horses in the non-SDF injury group could have been retired compulsorily at age 10 years rather than being retired due to an injury or other event. Another interesting comparison would have been that of horses with SDF injury to horses that retired due to any other injury that excluded age- and performance-related reasons for retirement.

Case-control

General description: A case-control study (Feinstein 1979; Ibrahim and Spitzer 1979; Miettinen 1985; Schulz and Grimes 2002) is an investigation in which the researcher selects a group of affected individuals (cases) and a comparison group (controls) to investigate factors associated with being a case. Controls are selected to represent the exposure distribution of the source population from which the cases developed (Miettinen 1985; Rothman and Greenland 1998, p 97) and there are many methods for selecting controls (Wacholder *et al.* 1992a; Grimes and Schulz 2005). Membership in the source population of the cases requires that the controls would have been captured as cases had they developed the outcome under study during the period of investigation. The source population of the cases can also be described as the study-base (Wacholder *et al.* 1992b; Miettinen 1999).

The study-base can be primary or secondary. A primary study-base is one in which the population of interest is defined first and then that population is examined to identify all cases. In order for controls to represent the exposure distribution of this population, they should be randomly sampled for the same time-period from which the cases were identified. A secondary study-base is one that is defined after the cases for study have been identified. In this situation it is not possible to randomly sample controls from the source population of the cases because it is typically not defined by a geographic region and time-period alone. A population-based case-control study is an example of an investigation with a primary base. Hospital-based case-control studies typically employ a secondary base approach. A secondary study-base is not easily defined in terms of an actual cohort or population and identification of an appropriate control group is an arduous task and fraught with the possibility of bias (Miettinen 1985; Wacholder *et al.* 1992c; Rothman and Greenland 1998, pp 93-114; Grimes and Schulz 2005).

Knowledge of outcome status (case and control) before the collection of exposure information from study subjects is the defining feature of a case-control study. The identification of factors predicting this outcome status must be the objective of the study. Cases of particular diseases that are examined for other outcomes (e.g. career earnings) should not be classified as employing this study design. Case-control studies are analytical (though the hypothesis could be vague) and can be defined as prospective or retrospective. A prospective (or incident) case-control study enrolls cases and controls in the present tense as cases develop and are identified. A retrospective case-control study investigates cases that had occurred prior to start of the study. A case-control study is formally defined by the following features: 1) the data are collected at the individual level; 2) the exposure groups for comparison are defined by observed characteristics not

derived by random assignment; 3) animals are sampled for study based on prior knowledge of outcome status; and 4) data could be collected for a single or multiple time-points. Information concerning exposure status, however, should be available for a time-point prior to development of the outcome of interest.

Case-control studies are considered the most efficient design for rare outcomes and relatively common exposures. The advantages are that they are relatively inexpensive to conduct, results can be obtained in a short time-period (particularly if a retrospective design is used), and it is possible to study multiple exposures if relevant data are available. Disadvantages include the fact that true risks cannot be calculated, they are limited to the investigation of a single outcome, they often rely on records and data that were not collected for the purpose of scientific investigation (so quality may be limited), selection of the appropriate control group(s) for the cases can be extremely difficult (and usually impossible to know if appropriate), and they are susceptible to many sources of bias.

True risks cannot be calculated from case-control studies because the prevalence of the disease (or other outcome of interest) is determined by the study design and not the disease prevalence in the population. For instance, a study that samples an equal number of controls for the identified cases will have 50% prevalence in the sample population fixed by design and this is not likely to be the prevalence in the source population. Although case-control studies are relatively easy to design they are a challenge to design well (Feinstein 1973; Schulz and Grimes 2002).

Hybrid designs: Modified case-control designs have been reported in the literature and include case-cohort, nested case-control, and case-crossover studies. Case-cohort and nested case-control studies can be performed when the population of interest is a well-defined cohort. All cases should be identified within the cohort during the study period and controls are either randomly sampled from all individuals at the start of the study (case-cohort) or sampled matched on time of occurrence for each case (nested case-control). Case-crossover studies record the exposure status of each case at the time of disease occurrence and the exposure status of the same individual related to another time prior to disease development. This design could be considered a specialised version of matching in a base case-control study or could be classified as a modification of a cohort or cross-sectional study depending upon the specifics of data collection.

Study example: A recent example of a case-control study is the paper by Boden *et al.* (2007) entitled "Risk factors for Thoroughbred racehorse fatality in jump starts in Victoria, Australia (1989–2004)." The objective of the study was "To identify risk factors for fatality of Thoroughbred racehorses in jump starts on all racecourses in Victoria, Australia between 1989 and 2004." Inherent in the authors' statement of objective is the expectation that some variables will be significant predictors of fatalities in this population of horses. The study evaluated a relatively large number of fatalities ($n = 191$) and randomly selected controls ($n = 2324$). Fatalities were significantly associated with racing career duration, number of starts prior to the race of interest, race-type (hurdle vs. steeplechase), calendar year and race-track location (country vs. city) using a multivariable logistic regression model with the inclusion of race-track as a random effect.

Thorough evaluation of potential sources of bias is critical for determining the quality of evidence provided by case-control studies. Of particular concern is selection bias that might result when selected cases do not represent the population (cases) that are the target of the evaluation (e.g. the selection of fatal cases when inferences are wanted about all types of injuries that occur during races). Inappropriate selection of cases can occur but is less common than errors related to inappropriate control group selection. The case-control study reported by Boden *et al.* (2007) employed a primary study-base defined as all jump starts in Victoria between 1st August 1989 and 31st July 2004. Reported fatalities during this time were assumed to represent all cases that occurred and exclusions were only performed when there were data discrepancies. Controls were selected randomly from a database for this same time-period and only identified cases were subsequently excluded.

The described methods and the primary study-base design indicate that selection bias was probably not a problem with this study. Information bias is a concern in retrospective case-control studies because data are typically not collected for the expressed purpose of scientific investigations and thus may have limitations both in the quality of data collection (e.g. incomplete detail of dietary histories for horses with colic and their controls) and the scope of data collection (e.g. absence of serum amyloid A concentrations among horses with colic and their controls). Data for the study by Boden *et al.* (2007) were extracted from 3 major databases and are likely to be of high-quality, thereby minimising the likelihood of errors in classifying outcome status. The authors performed appropriate statistical methods to control for factors under study, but it is impossible to know if other factors not available in the databases might have confounded measured associations.

The authors mention in the discussion that their model was unable to assess the effects of going and weather on measured associations, which have previously been associated with falls. Whenever possible, case-control studies should collect data for previously identified risk factors to account for confounding variables; when previously published confounders are not accounted for, study results should be interpreted with caution. Generally, the extent of the impact of ignored confounders will be commensurate with the magnitude of the association of the ignored confounder (e.g. weather) and the outcome (e.g. injury due to a fall). Clearly, it is not always possible to collect data on all variables previously associated with the outcome of interest and failure to account for a known confounder does not necessarily vitiate the value of a study.

Cross-sectional

General description: A cross-sectional study (Kelsey 1996, pp 244-257) is an investigation in which the researcher selects for study a group of individuals that is often defined by membership in a target population at a particular point in time. These studies are typically descriptive in nature but they could also be designed to test specific hypotheses and therefore be analytical. They typically employ prospective data collection but retrospective designs are theoretical alternatives to retrospective case-control studies. The defining aspect of this study design is that the act of measuring outcome and exposure status is performed at the same time-point. The more usual definition is that the outcome and exposure relate to the same time-point (Rothman 2002, pp 89-91;

Dohoo *et al.* 2003, pp 144-147). However, it is possible to survey a group of individuals for the assessment of outcome and exposure with the same instrument (e.g. questionnaire) and include questions related to past exposures. A cross-sectional study is formally defined by the following: 1) data are collected at the individual level; 2) exposure groups for comparison are defined by observed characteristics not derived by random assignment; 3) animals are typically sampled for study based on membership in a certain population (not required); and 4) data concerning outcome and exposure are either collected related to the same time-point or data are collected by investigators at the same time.

Information provided by cross-sectional studies is the prevalence of exposure variables and outcome and characteristics of affected and unaffected individuals. Advantages of these studies are that they are relatively inexpensive and easy to design and implement, and results are easily analysed. They are most appropriate for evaluating exposure-outcome relationships for factors that are fixed (e.g. breed, sex, genotype). Disadvantages include that there is no temporal (longitudinal) component so they cannot be used to estimate incidence and it is often difficult to assess a causal relationship.

Study example: A recent example of a cross-sectional study is an investigation by Hotchkiss *et al.* (2007a) entitled "A survey of horse owners in Great Britain regarding horses in their care. Part 2: Risk factors for recurrent airway obstruction." The objectives of this study were "To estimate the prevalence of recurrent airway obstruction (RAO) in the general horse population of Great Britain and to investigate possible risk factors for RAO associated with management or early life." The study employed a mailed questionnaire sent to 1431 clients from 14 veterinary practices; 873 were returned and available for statistical analysis. The prevalence of RAO was estimated to be 14% and identified risk factors were increasing age, residing in an urbanised environment, exposure to hay, and a history of a respiratory infection early in life. Although the authors neither provided a rationale for the sample size nor described the proportional sampling at each stage of the design, these descriptions were included in a previous study by these authors (Hotchkiss *et al.* 2007b). Random-effects multivariable logistic regression was used to account for the clustered sampling design.

Potential sources of systematic error in this study included selection, information, and confounding biases. Random selection of veterinary practices and clients within practices should have reduced the possibility of selection bias. However, responder bias occurs when the horses owned by people that return a completed questionnaire are systematically different from those that did not return a questionnaire and can be considered a type of selection bias. High response proportions are typically assumed to be less biased than lower proportions (Jones 1996; Roush 1998), though this is not always the case (Stang and Jockel 2004). The authors reported a reasonable response proportion, but if possible it is beneficial when reporting studies to present descriptive data related to the comparability of responders and nonresponders in terms of characteristics that might be available in client records or through a subset validation.

The authors discussed the possibility of misclassification and recall bias, which are types of information bias. The authors employed a standardised assessment tool (risk screening questionnaire) and handled data acquisition consistently for all selected owners. Potential sources of information bias existed due

to the nature of the study but this does not suggest errors of the study *per se*. The statistical approach of the authors was appropriate to investigate and adjust for confounding that might have resulted from measured variables.

The inability to determine whether the exposures under investigation preceded the outcome is a limitation of cross-sectional designs in general. The authors recognised this possibility and discussed the specific example that wetting hay before feeding seemed to increase the odds of RAO in horses. However, in this cross-sectional study, feeding wetted hay was probably a result of horses being diagnosed with RAO rather than being a cause of RAO. This study appears to have been well-designed and executed with the highest degree of scientific rigour. Potential limitations in assuming causal associations are due to issues of the base study design rather than errors of implementation. Investigators must weigh the limitations of various study designs against cost and feasibility of other approaches. The primary objective of estimating prevalence in this population of horses could not have been achieved easily by other base study designs.

Case series

General description: Case series (Grimes and Schulz 2002d) are descriptive observational studies used to generate hypotheses concerning predictors of disease or recovery that can be performed retrospectively or prospectively. However, a case series with sequential data collection (covering multiple time-points) is a type of cohort study where membership is defined by the case or procedure of interest.

The study design is defined by the factors: 1) data are collected at the individual animal level; 2) animals can be grouped for *post hoc* statistical comparisons based on exposures but assignment to groups is not performed by the investigators; 3) animals are sampled to represent the disease or therapeutic procedure of interest; and 4) information can be collected prospectively or from retrospective sources related either to a single or multiple time-points. Case series are relatively easy to perform and are inexpensive, especially when performed using retrospective cases/data. A disadvantage of this design is the inability to test specific hypotheses.

Study example: A recent example of a case series design is the study by Levine and Richardson (2007) entitled "Clinical use of the locking compression plate (LCP) in horses: a retrospective study of 31 cases (2004–2006). The objective of this study was "To describe a series of clinical cases in which the LCP was used for fracture stabilisation or arthrodesis." Twenty-seven (87%) of the horses were discharged from the hospital and recognised complications included incisional infection (32%), implant infection (19%), implant loosening/breakage (22%), contralateral limb laminitis (16%), colic (3%) and diarrhoea (3%). Twenty-five (81%) were sound for their intended purpose at long-term follow-up (minimum of 6 weeks). Statistical procedures were not employed as part of this study because no comparisons were desired. The estimates for discharge proportion and recognised complications could theoretically be biased, but inferences related to other repair methods were not made. It is possible that the selected cases do not represent all cases in which LCP might be utilised (selection bias) and information bias might have affected the collection of data related to complications. Confounding could

not be a source of error because no statistical associations were examined. The authors thoroughly described the cases, and the descriptive nature of the study precluded concerns related to systematic or random errors.

Results and causality

The primary purpose of epidemiological investigations is to learn about causal mechanisms related to disease incidence and to identify factors for therapy and prevention. Presentation of epidemiological findings should include a point-estimate for the measure of association (e.g. odds ratio), measure of the random error associated with the point estimate, and the P value for the statistical test of the measured association compared to some null value. Random error can be quantified using confidence intervals or reporting the standard error. The definition and appropriate use of P values is often misunderstood (Freeman 1993; Goodman 1993; Brennan and Croft 1994; Sterne and Davey Smith 2001; Gellerstedt 2002; Sterne 2002; Ioannidis 2005) and significant associations are not necessarily causal associations.

Identification of factors associated with disease incidence, resolution and prevention is important when the associations are causal. Theories related to causation (Rothman 1976; Krieger 1994) and causal inference (Weed 1986; Greenland 1990; Rothman and Greenland 2005; Lash 2007) have been described that emphasise the fact that there is no simple means to determine if an association is truly causal. Guidelines and criteria for determining causal associations have been published (Parascandola *et al.* 2006) and can serve as a useful starting point; however, no simple checklist alone is sufficient. The decision to accept a relationship as causal should consider: 1) the chance that the observed association (or more extreme observations) occurred due to random variation (measure of statistical significance); 2) the possibility that the 2 factors are intrinsically related (e.g. day follows night); 3) the possibility that the measured association was biased; and 4) causal guidelines, such as Bradford Hill's modification (Hill 1965) of the Surgeon General's report concerning smoking and health (Anon 1964). The first 3 considerations are germane to results of a single study, while Hill's guidelines relate to the study itself in addition to cumulative evidence and scientific reasoning.

The guidelines of Hill include: 1) consideration of time-sequence (i.e. purported cause must precede effect); 2) strength (magnitude) of association; 3) evidence of a dose-response relationship; 4) consistency of findings among studies; 5) biological plausibility of the observed association; 6) specificity of the association (i.e. the purported cause is not associated with multiple effects/outcomes); 7) analogy of the association (i.e. the association is similar to a condition with a known causal pathway); and 8) experimental evidence supporting the observed association. Epidemiologists argue the usefulness of

causal guidelines and more information concerning these guidelines can be found elsewhere (Hofler 2005, 2006; Rothman and Greenland 2005; Phillips and Goodman 2006). The only listed guideline that is a true criterion for causation is the time-sequence of events (temporality); by definition, a cause must precede the event. In some instances (e.g. the study of chronic diseases with indistinct onset), clearly establishing the temporal association of exposure and disease can be difficult.

Refutationism revisited

Based on the refutationist view of scientific discovery, the motivation of investigators should be to refute their scientific hypothesis. Rejection of a statistical null hypothesis can be taken as evidence that collected data are not consistent with that stated hypothesis. However, a statistical null hypothesis (no difference between groups) is typically the converse of the research hypothesis (there is a difference between groups). The philosophical concern is that the typical scientific approach is to design a study to reject a statistical null hypothesis that is often opposite to the research hypothesis. Therefore, the research hypothesis is typically not formally tested. Furthermore, statistical tests are based on allowable *type I* and *type II* error limits (Neyman and Pearson 1928, 1933; Jones *et al.* 2003) that highlight uncertainty and explain why rejection of a null hypothesis is not the same as refutation of the same hypothesis (Table 4). This emphasises the concern that has been previously discussed (Pearce 1990) that only hypotheses stated as 'never occurs' or 'does not exist' can be refuted by a single study. The hypothesis that all swans are white can be refuted by the observation of a single swan of a different colour. However, the hypothesis that black swans are smaller in size than their white counterparts cannot be refuted by the observation of a single large black swan. The refutationist approach should be applied to the process of scientific discovery; however, most hypotheses of medical and epidemiological importance will not be refutable because questions related to causality cannot be formulated in such a manner and empirical science is based on observations that might be in error. Therefore, the refutationist approach is important for the design of a study but is not particularly useful for the interpretation of findings.

Conclusions

The study design employed to measure the association between a factor and an outcome has a great impact on the quality of evidence for determining causal associations. Descriptive studies do not provide evidence for making causal inferences that is as strong as that of studies designed to test specific hypotheses. Studies that are not able to determine that the purported cause preceded the outcome also will not provide strong evidence for causality. Since it is never possible to determine the accuracy of a study, they must be judged by assessing the likelihood of bias in the study design and analysis.

All studies should be assumed to contain bias and it is imperative for investigators to document for readers the strategies employed to minimise the detrimental effects of bias that might exist in their study. By comprehensively describing the aforementioned information, investigators reduce the onus on readers with respect to assessing the quality of the clinical evidence in their report. Experimental studies are least likely to

TABLE 4: Statistical errors associated with hypothesis testing. The probability of making a *Type I* and *Type II* error are typically referred to as alpha and beta, respectively

		True state of nature	
		H ₀ is true	H ₀ is false (H _A true)
Statistical result	Reject H ₀	Alpha	Correct result
	Do not reject H ₀	Correct result	Beta

H₀ = statistical null hypothesis; H_A = alternative statistical hypothesis.

contain biases when they are implemented appropriately but they are less commonly performed than observational studies. The defining characteristics of commonly employed observational studies are important to understand because they are frequently misrepresented in the literature and the base design is important for the assessment of bias.

Experimental studies (RCTs) are generally considered to provide the highest quality of evidence for evaluating clinical procedures because when performed properly they are less likely to contain systematic errors. Results of a study can be accurate (close to the true values) irrespective of systematic errors and bias. However, on average results are more likely to be close to the true value in the absence of systematic error. Results from high-quality RCTs provide stronger evidence for determining the true state of nature but the role of random error prevents results from a single study from being sufficient for proving causality. Also, scientific reasoning plays a major role in the qualitative assessment of study validity and currently unknown sources of systematic error could unknowingly bias results.

Cohort studies can be prospective or retrospective and the defining characteristic is that a group is defined based on a shared characteristic and followed over time for the measurement of an outcome. These studies can be expensive and labour intensive. They are considered to be the most efficient design for the investigation of relatively frequent outcomes and rare exposures if exposure status is used for the definition of cohorts (i.e. exposed cohort and unexposed cohort). A prospective cohort design should be considered when other designs have yielded contradictory results and an experimental study would not be possible or feasible to perform.

Cross-sectional studies are often employed by sampling a group with a shared characteristic. This targeted group could be considered a cohort because of this identifying feature. What distinguishes this study from a cohort study is that the targeted group is not followed over time to measure the incidence of an outcome. Data are collected from the sampled subjects only a single time or all collected data relate to the same point in time. A single questionnaire could be used to collect data related to multiple time points but what distinguishes this from a retrospective cohort is that the cohort is defined based on a characteristic in the present and the data correspond to past exposures.

Case-control studies are generally considered the most efficient design to study rare diseases and common exposures. They can be performed prospectively if cases are captured by the study as they occur; however, more frequently they rely on retrospective data from cases that occurred prior to study initiation. The defining feature of this study design is that knowledge related to the outcome is known prior to collection of data concerning potential predictors of the outcome. The study of cases of a disease compared to noncases of that disease is not sufficient for the classification of this design. Case status must be the outcome of interest.

Case series are descriptive studies that could be used to generate hypotheses. They can be performed prospectively or be based on retrospective data. They are generally performed when few data exist related to the disease or procedure of interest. They are especially useful for new or exceedingly rare outcomes or therapies.

Equine health can be improved with the incorporation of results derived from well-designed and implemented

epidemiological studies. It is the obligation of investigators to implement and adequately describe procedures to reduce the impact of bias on measured results. It is important to emphasise our roles as unbiased scientists over our inner advocates that desire to collect only evidence that supports our predetermined beliefs. Given that the best evidence for clinical practice will likely be derived from patient-centred observational epidemiological studies for the foreseeable future, it is imperative that equine veterinarians become familiar with epidemiological study designs and their respective strengths and limitations. There is also an onus on investigators to describe potential biases and, when possible, to quantify the magnitude of these potential biases. Understanding epidemiological study design and the principles of interpreting these studies is germane to all informed equine practitioners and equine clinical investigators.

References

- Altman, D.G., Schulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gotzsche, P.C. and Lang, T. (2001) The revised CONSORT statement for reporting randomised trials: explanation and elaboration. *Ann. intern. Med.* **134**, 663-694.
- Anon (1964) Smoking and health: report of the Advisory Committee to the Surgeon General of the Public Health Service, US Department of Health and Welfare, Washington, DC.
- Archer, D.C. and Proudman, C.J. (2006) Epidemiological clues to preventing colic. *Vet. J.* **172**, 29-39.
- Balkau, B. and Eschwege, E. (1995) Risk factors and their identification second part: study designs for identification of risk factors. *Diabete. Metab.* **21**, 225-231.
- Baxter, H. (2001) Understanding research: 1. Methodologies used to answer research questions. *J. Wound. Care.* **10**, 269-272.
- Benson, K. and Hartz, A.J. (2000) A comparison of observational studies and randomised, controlled trials. *N. Engl. J. Med.* **342**, 1878-1886.
- Birkett, N.J. (1992) Effect of nondifferential misclassification on estimates of odds ratios with multiple levels of exposure. *Am. J. Epidemiol.* **136**, 356-362.
- Boden, L.A., Anderson, G.A., Charles, J.A., Morgan, K.L., Morton, J.M., Parkin, T.D., Clarke, A.F. and Slocombe, R.F. (2007) Risk factors for Thoroughbred racehorse fatality in jump starts in Victoria, Australia (1989-2004). *Equine vet. J.* **39**, 422-428.
- Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Moher, D., Rennie, D., de Vet, H.C. and Lijmer, J.G. (2003) The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann. intern. Med.* **138**, W1-12.
- Boutron, I., Moher, D., Altman, D.G., Schulz, K.F. and Ravaud, P. (2008) Extending the CONSORT statement to randomised trials of nonpharmacologic treatment: explanation and elaboration. *Ann. intern. Med.* **148**, 295-309.
- Brennan, P. and Croft, P. (1994) Interpreting the results of observational research: chance is not such a fine thing. *Br. med. J.* **309**, 727-730.
- Brenner, H. (1991) Re: "Does nondifferential misclassification of exposure always bias a true effect toward the null value?" *Am. J. Epidemiol.* **134**, 2.
- Brenner, H. and Loomis, D. (1994) Varied forms of bias due to nondifferential error in measuring exposure. *Epidemiol.* **5**, 510-517.
- Buck, C. (1975) Popper's philosophy for epidemiologists. *Int. J. Epidemiol.* **4**, 159-168.
- Christenfeld, N.J., Sloan, R.P., Carroll, D. and Greenland, S. (2004) Risk factors, confounding, and the illusion of statistical control. *Psychosom. Med.* **66**, 868-875.
- Chuang, T.Y. and Reizner, G.T. (1993) Dermatopidemiology. Part I: Epidemiologic methods. *Int. J. Dermatol.* **32**, 251-256.
- Concato, J., Shah, N. and Horwitz, R.I. (2000) Randomised, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med.* **342**, 1887-1892.
- Coomer, R.P., Mair, T.S., Edwards, G.B. and Proudman, C.J. (2007) Do subcutaneous sutures increase risk of laparotomy wound suppuration? *Equine vet. J.* **39**, 396-399.
- Copeland, K.T., Checkoway, H., McMichael, A.J. and Holbrook, R.H. (1977) Bias due to misclassification in the estimation of relative risk. *Am. J. Epidemiol.* **105**, 488-495.

- Des Jarlais, D.C., Lyles, C. and Crepaz, N. (2004) Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *Am. J. Public Health* **94**, 361-366.
- Dickey, J.O., Marcus, S.L., de Viron, O. and Fukumori, I. (2002) Recent Earth oblateness variations: Unraveling climate and postglacial rebound effects. *Science* **298**, 1975-1977.
- Dohoo, I.R., Martin, S.W. and Stryhn, H. (2003) *Veterinary Epidemiologic Research*, AVC Inc., Charlottetown. 144-147.
- Feinstein, A.R. (1973) Clinical biostatistics. XX. The epidemiologic trohoc, the ablative risk ratio, and "retrospective" research. *Clin. Pharmacol. Ther.* **14**, 291-307.
- Feinstein, A.R. (1979) Methodologic problems and standards in case-control research. *J. chronic Dis.* **32**, 35-41.
- Feinstein, A.R. (1988) Scientific standards in epidemiologic studies of the menace of daily life. *Science* **242**, 1257-1263.
- Flegel, K.M., Keyl, P.M. and Nieto, F.J. (1991) Differential misclassification arising from nondifferential errors in exposure measurement. *Am. J. Epidemiol.* **134**, 1233-1244.
- Fosgate, G.T. (2006) Non-differential measurement error does not always bias diagnostic likelihood ratios towards the null. *Emerg. Themes Epidemiol.* **3**, 7.
- Freeman, P.R. (1993) The role of p-values in analysing trial results. *Stat. Med.* **12**, 1443-1452; discussion 1453-1448.
- Freudenheim, J.L. (1999) Study design and hypothesis testing: issues in the evaluation of evidence from research in nutritional epidemiology. *Am. J. clin. Nutr.* **69**, 1315S-1321S.
- Gellerstedt, M. (2002) Statistical issues - significantly important in medical research. *Allergy* **57**, 76-82.
- Gladen, B. and Rogan, W.J. (1979) Misclassification and the design of environmental studies. *Am. J. Epidemiol.* **109**, 607-616.
- Goodman, S.N. (1993) P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am. J. Epidemiol.* **137**, 485-496.
- Greenland, S. (1990) Randomization, statistics, and causal inference. *Epidemiol.* **1**, 421-429.
- Greenland, S., Gago-Dominguez, M. and Castela, J.E. (2004) The value of risk-factor ("black-box") epidemiology. *Epidemiol.* **15**, 529-535.
- Greenland, S. and Morgenstern, H. (2001) Confounding in health research. *Ann. Rev. Public Health* **22**, 189-212.
- Grimes, D.A. and Schulz, K.F. (2002a) Bias and causal associations in observational research. *Lancet* **359**, 248-252.
- Grimes, D.A. and Schulz, K.F. (2002b) Cohort studies: marching towards outcomes. *Lancet* **359**, 341-345.
- Grimes, D.A. and Schulz, K.F. (2002c) Descriptive studies: what they can and cannot do. *Lancet* **359**, 145-149.
- Grimes, D.A. and Schulz, K.F. (2002d) An overview of clinical research: the lay of the land. *Lancet* **359**, 57-61.
- Grimes, D.A. and Schulz, K.F. (2005) Compared to what? Finding controls for case-control studies. *Lancet* **365**, 1429-1433.
- Harris, S.A. (2000) Epidemiology: theory, study design, and planning for education. *J. Contin. Educ. Health Prof.* **20**, 133-145.
- Hill, A.B. (1965) The environment and disease: Association or causation? *Proc. R. Soc. Med.* **58**, 295-300.
- Hofler, M. (2005) The Bradford Hill considerations on causality: a counterfactual perspective. *Emerg. Themes Epidemiol.* **2**, 11.
- Hofler, M. (2006) Getting causal considerations back on the right track. *Emerg. Themes Epidemiol.* **3**, 8.
- Horwitz, R.I. (1987) The experimental paradigm and observational studies of cause-effect relationships in clinical medicine. *J. chronic Dis.* **40**, 91-99.
- Hotchkiss, J.W., Reid, S.W. and Christley, R.M. (2007a) A survey of horse owners in Great Britain regarding horses in their care. Part 2: Risk factors for recurrent airway obstruction. *Equine vet. J.* **39**, 301-308.
- Hotchkiss, J.W., Reid, S.W. and Christley, R.M. (2007b) A survey of horse owners in Great Britain regarding horses in their care. Part 1: Horse demographic characteristics and management. *Equine vet. J.* **39**, 294-300.
- Ibrahim, M.A. and Spitzer, W.O. (1979) The case control study: the problem and the prospect. *J. chronic Dis.* **32**, 139-144.
- Ioannidis, J.P. (2005) Why most published research findings are false. *PLoS Med.* **2**, e124.
- Ioannidis, J.P., Evans, S.J., Gotzsche, P.C., O'Neill, R.T., Altman, D.G., Schulz, K. and Moher, D. (2004) Better reporting of harms in randomised trials: an extension of the CONSORT statement. *Ann. intern. Med.* **141**, 781-788.
- Jones, J. (1996) The effects of non-response on statistical inference. *J. Health Soc. Policy* **8**, 49-62.
- Jones, S.R., Carley, S. and Harrison, M. (2003) An introduction to power and sample size estimation. *Emerg. Med. J.* **20**, 453-458.
- Jurek, A.M., Greenland, S. and Maldonado, G. (2008) How far from non-differential does exposure or disease misclassification have to be to bias measures of association away from the null? *Int. J. Epidemiol.* **37**, 382-385.
- Kelsey, J.L. (1996) *Methods in Observational Epidemiology*, 2nd edn., Oxford University Press, Oxford. pp 3, 244-257.
- King, L.S. (1983) Medicine in the USA: historical vignettes. VIII. Germ theory and its influence. *J. Am. med. Ass.* **249**, 794-798.
- Kramer, M.S. and Boivin, J.F. (1987) Toward an "unconfounded" classification of epidemiologic research design. *J. chronic Dis.* **40**, 683-688.
- Krieger, N. (1994) Epidemiology and the web of causation: has anyone seen the spider? *Soc. Sci. Med.* **39**, 887-903.
- Lam, K.H., Parkin, T.D., Riggs, C.M. and Morgan, K.L. (2007) Descriptive analysis of retirement of Thoroughbred racehorses due to tendon injuries at the Hong Kong Jockey Club (1992-2004). *Equine vet. J.* **39**, 143-148.
- Lash, T.L. (2007) Heuristic thinking and inference from observational epidemiology. *Epidemiol.* **18**, 67-72.
- Last, J.M. and International Epidemiological Association. (2001) *A Dictionary of Epidemiology*, 4th edn., Oxford University Press, New York.
- Levine, D.G. and Richardson, D.W. (2007) Clinical use of the locking compression plate (LCP) in horses: a retrospective study of 31 cases (2004-2006). *Equine vet. J.* **39**, 401-406.
- Marshall, T. (2004) What is a case-control study? *Int. J. Epidemiol.* **33**, 612-613.
- McKee, M., Britton, A., Black, N., McPherson, K., Sanderson, C. and Bain, C. (1999) Methods in health services research. Interpreting the evidence: choosing between randomised and non-randomised studies. *Br. med. J.* **319**, 312-315.
- Mertens, T.E. (1993) Estimating the effects of misclassification. *Lancet* **342**, 418-421.
- Miettinen, O.S. (1985) The "case-control" study: valid selection of subjects. *J. chronic Dis.* **38**, 543-548.
- Miettinen, O.S. (1999) Etiologic research: needed revisions of concepts and principles. *Scand J. Work Environ. Health* **25**, 484-490.
- Miettinen, O.S. and Cook, E.F. (1981) Confounding: essence and detection. *Am. J. Epidemiol.* **114**, 593-603.
- Moher, D., Schulz, K.F. and Altman, D.G. (2001) The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* **357**, 1191-1194.
- Neyman, J. and Pearson, E.S. (1928) On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika* **20A**, 175-240.
- Neyman, J. and Pearson, E.S. (1933) On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **231**, 289-337.
- Parascandola, M., Weed, D.L. and Dasgupta, A. (2006) Two Surgeon General's reports on smoking and cancer: a historical investigation of the practice of causal inference. *Emerg. Themes Epidemiol.* **3**, 1.
- Parkin, T.D. (2008) Epidemiology of racetrack injuries in racehorses. *Vet. Clin. N. Am.: Equine Pract.* **24**, 1-19.
- Pearce, N. (1990) White swans, black ravens, and lame ducks: necessary and sufficient causes in epidemiology. *Epidemiol.* **1**, 47-50.
- Phillips, C.V. and Goodman, K.J. (2006) Causal criteria and counterfactuals; nothing more (or less) than scientific common sense. *Emerg. Themes Epidemiol.* **3**, 5.
- Piaggio, G., Elbourne, D.R., Altman, D.G., Pocock, S.J. and Evans, S.J. (2006) Reporting of noninferiority and equivalence randomised trials: an extension of the CONSORT statement. *J. Am. med. Ass.* **295**, 1152-1160.
- Pinchbeck, G.L., Clegg, P.D., Proudman, C.J., Morgan, K.L. and French, N.P. (2004) A prospective cohort study to investigate risk factors for horse falls in UK hurdle and steeplechase racing. *Equine vet. J.* **36**, 595-601.
- Poole, C. and Rothman, K.J. (1998) Our conscientious objection to the epidemiology wars. *J. Epidemiol. Community Health* **52**, 613-614.
- Popper, K.R. (1963) Science: Problems, aims, responsibilities. *Fed. Proc.* **22**, 961-972.
- Rothman, K.J. (1976) Causes. *Am. J. Epidemiol.* **104**, 587-592.
- Rothman, K.J. (2002) *Epidemiology: an introduction*, Oxford University Press, Oxford. pp 89-91, 94-95.
- Rothman, K.J. and Greenland, S. (1998) *Modern Epidemiology*, 2nd edn., Lippincott-Raven, Philadelphia. pp 93-114.

- Rothman, K.J. and Greenland, S. (2005) Causation and causal inference in epidemiology. *Am. J. Public Health* **95**, Suppl. 1, S144-150.
- Rothwell, P.M. (2005) External validity of randomised controlled trials: "to whom do the results of this trial apply?" *Lancet* **365**, 82-93.
- Roush, J.K. (1998) Nonselection and nonresponse bias in clinical research. *J. Am. vet. med. Ass.* **213**, 1270-1273.
- Rubinfeld, G.D. (1995) Study design: statistical and methodologic considerations. *Respir. Care* **40**, 980-986.
- Sackett, D.L. (1979) Bias in analytic research. *J. chronic Dis.* **32**, 51-63.
- Schulz, K.F. and Grimes, D.A. (2002) Case-control studies: research in reverse. *Lancet* **359**, 431-434.
- Skrabaneck, P. (1994) The emptiness of the black box. *Epidemiol.* **5**, 553-555.
- Stang, A. and Jockel, K.H. (2004) Studies with low response proportions may be less biased than studies with high response proportions. *Am. J. Epidemiol.* **159**, 204-210.
- Stephenson, J.M. and Babiker, A. (2000) Overview of study design in clinical epidemiology. *Sex Transm. Infect.* **76**, 244-247.
- Sterne, J.A. (2002) Teaching hypothesis tests - time for significant change? *Stat. Med.* **21**, 985-994; discussion 995-999, 1001.
- Sterne, J.A. and Davey Smith, G. (2001) Sifting the evidence - what's wrong with significance tests? *Br. med. J.* **322**, 226-231.
- Szklo, M. and Nieto, F.J. (2000) *Epidemiology: Beyond the Basics*, Aspen, Gaithersburg. pp 125-126.
- Vandenbroucke, J.P., von Elm, E., Altman, D.G., Gotzsche, P.C., Mulrow, C.D., Pocock, S.J., Poole, C., Schlesselman, J.J. and Egger, M. (2007) Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Ann. intern. Med.* **147**, W163-194.
- Wacholder, S., Silverman, D.T., McLaughlin, J.K. and Mandel, J.S. (1992a) Selection of controls in case-control studies. II. Types of controls. *Am. J. Epidemiol.* **135**, 1029-1041.
- Wacholder, S., McLaughlin, J.K., Silverman, D.T. and Mandel, J.S. (1992b) Selection of controls in case-control studies. I. Principles. *Am. J. Epidemiol.* **135**, 1019-1028.
- Wacholder, S., Silverman, D.T., McLaughlin, J.K. and Mandel, J.S. (1992c) Selection of controls in case-control studies. III. Design options. *Am. J. Epidemiol.* **135**, 1042-1050.
- Weed, D.L. (1985) An epidemiological application of Popper's method. *J. Epidemiol. Community Health* **39**, 277-285.
- Weed, D.L. (1986) On the logic of causal inference. *Am. J. Epidemiol.* **123**, 965-979.
- Weed, D.L. (1998) Beyond black box epidemiology. *Am. J. Public Health* **88**, 12-14.
- Young, J.O. (2001) A defense of the coherence theory of truth. *J. Philosophical Res.* **26**, 13.

Appendix

Glossary of terms

Accuracy: The extent to which a measurement represents the true value of the quantity being measured.

Association: A relationship (dependency) between 2 or more factors.

Association, causal: A relationship between 2 factors that is due to one factor being a cause of the other.

Association, significant: A relationship between 2 or more factors that is found to be statistically different from a null value at a prescribed level of random error.

Bias: A systematic error in the study design, data collection, or analysis that causes findings to deviate from the truth.

Bias, confounding: A systematic error caused by failure to account for the effect of another factor.

Bias, information: A systematic error in data collection or recording that causes the quality of information to be different between comparison groups.

Bias, selection: A systematic error in the inclusion of study participants resulting in a study population that is not representative of the target population of interest with respect to the study purpose.

Case-control study: An epidemiological study in which the sampling of participants is dependent upon prior knowledge of disease status (outcome).

Case series study: An epidemiological study that describes a group of subjects with a particular condition of interest that does not include a comparison group.

Cause: A factor that was necessary for an event (outcome) to occur at the time that it did.

Cross-sectional study: An epidemiological study where all data relate to a single point in time or were collected at the same time (e.g. a single questionnaire administration).

Cohort: Any group of subjects that can be defined by a shared characteristic.

Cohort study: An epidemiological study that follows a group or groups over time to measure the incidence of an outcome of interest.

Confidence interval: A range of possible true values of a parameter that is reasonably consistent with the collected data based on expected random error.

Confounding: The distortion of a measured association between 2 factors caused by some other variable(s).

Effect: The direct consequence of a causal factor on a measured outcome.

Effect modification: The variation of a measured association between 2 factors among categories of a different factor.

Epidemiology: The study of the distribution and determinants of health states in populations or the body of knowledge resulting from such studies.

Epidemiology, analytical: The collection of theories and methods employed for the study of the distribution and determinants of health states in populations or their development.

Epidemiology, applied: The use of recognised methods for the study of the distribution and determinants of health states in populations.

Epidemiology, descriptive: The use of recognised methods to describe the distribution of health states in populations.

Error, random: Variation due to imprecision in a measuring instrument that cannot be predicted by other factors.

Error, systematic: Variation that is persistent, has a nonzero mean, and can be attributed to the effect of other factors.

Error, Type I: The probability of rejecting the null hypothesis when it is true. The amount of error is often referred to as alpha (α).

Error, Type II: The probability of not rejecting the null hypothesis when the null hypothesis is false and a particular alternative hypothesis is true. The amount of error is often referred to as beta (β).

Evidence: That which tends to prove or disprove something; knowledge forming the basis for a belief, conclusion, or judgment.

Exposure: A factor that is investigated as a potential cause of an outcome.

Hypothesis: A testable conjecture that could be a proposed association between 2 factors.

Hypothesis, alternative: The hypothesis that includes all possible values not included in the null hypothesis. This hypothesis is accepted upon rejection of the null.

Hypothesis, null: The assumed relationship between 2 factors that is used to perform statistical testing. The typical choice for the null hypothesis is no effect or no difference between groups.

Incidence: The amount of newly developed cases of the outcome over a specified time period. It can be measured as either a proportion or a rate.

Odds: The ratio of the probability that an event will occur divided by the probability that the event will not occur.

Odds ratio: The ratio of 2 odds that is used to quantify the relationship between 2 factors.

Outcome: The factor that is investigated to determine relationships with potential causes. Common outcomes are disease or infection states.

P value: The probability of observing the collected data or more extreme when the null hypothesis is true.

Parameter: A measurable characteristic of a population.

Population: The entire collection of units.

Population, study: The collection of units in which data were collected.

Population, target: The collection of units from which the study population was selected.

Precision: The inverse of the variation of a set of measurements.

Prevalence: The proportion of cases of the outcome within a sampled population at a given point in time or over a defined time period.

Probability: The relative possibility that an event will occur, as expressed by the proportion of the number of occurrences to the total number that could have occurred.

Proportion: A ratio in which the numerator is a subset of the denominator.

Prospective: Occurring in the present tense.

Randomised control trial (RCT): An experimental epidemiological study in which the exposure of interest is randomly assigned to study participants. Study participants are frequently cases of disease and the effect of treatment or therapy on disease resolution is measured.

Rate: A ratio of the number of newly developed outcomes within a specified study population divided by the total amount of outcome-free time within the population.

Rate ratio: A ratio of 2 rates that is used to quantify the relationship between 2 factors.

Ratio: A general value that is formed by a numerator divided by a denominator. Rates, proportions and odds are specific ratios.

Refutationism: The philosophy of science put forth by K.R. Popper that promotes the idea that scientific investigations should attempt to falsify, or refute, hypotheses rather than rely on inductive logic in effort to support them.

Retrospective: Related to past events.

Risk: The probability that an event will occur within a specified population over a certain period of time.

Risk ratio: A ratio of 2 risks that is used to quantify the relationship between 2 factors.

Study, analytic: An epidemiological study that is performed to test a research hypothesis explicitly defined prior to data collection.

Study, descriptive: An epidemiological study that is performed to investigate an outcome without a specified research hypothesis stated prior to data collection.

Study design: The architecture or blueprint including all details necessary to perform a study.

Validity: The degree to which the system employed to make measurements is free from sources of systematic error.