Study design synopsis: Bias can cast a dark shadow over studies

Geoffrey Fosgate

Department of Production Animal Studies, Faculty of Veterinary Science, University of Pretoria, Onderstepoort, South Africa

Correspondence to Geoffrey Fosgate, Department of Production Animal Studies, Faculty of Veterinary Science, University of Pretoria, Onderstepoort, South Africa.
Email: Geoffrey.fosgate@up.ac.za

Abstract

The study of free-living populations is important to generate knowledge related to the epidemiology of disease and other health outcomes. These studies are unable to provide the same level of control as is possible in laboratory studies and thus are susceptible to certain errors. The primary categories of study errors are random and systematic. Random errors cause imprecision and can be quantified using statistical methods including the calculation of confidence intervals. Systematic errors cause bias, which is typically difficult to quantify within the context of an individual study. The three main categories of systematic errors are selection, information, and confounding bias. Selection bias occurs when enrolled animals are not representative of the target population of interest in respect to characteristics important to the primary study objective. Information bias occurs when data collected from enrolled animals deviates from the true value. Information bias is most damaging when errors vary among comparison groups. Both selection and information bias are prevented through the application of good study design procedures. Researchers should select study animals after careful consideration of the primary study objective and desired target population. Investigators can reduce information bias through standardised data collection procedures and the use of blinding. Confounding bias occurs when the measured association between a predictor and an outcome ignores the influential effect of an additional variable. Confounding is common and analysts must implement the appropriate statistical adjustments to reduce the associated bias. All studies will have some errors and biased data with high precision are the most damaging to the validity of study conclusions. Authors can facilitate the critical evaluation of their research by providing text related to the limitations and potential sources of bias within the discussion section of their manuscripts.

Keywords: horse; confounding; epidemiology; random error; systematic error

1 INTRODUCTION

The number of published scientific manuscripts is increasing substantially with time[1, 2]and there is a belief that this increase is associated with a reduction in overall research quality.[3] The increase in publications makes it difficult for researchers to stay abreast of the current literature due to extensive time demands within academia.[4] It is therefore necessary for scientists to develop the required tools for the critical evaluation of literature in an effort to allocate time to high-quality research while minimising time devoted to lower-quality work.

To create high-quality research, researchers should design studies in a manner consistent with the current state of knowledge in the field and follow established standards[5, 6] to improve reporting and reduce the influence of errors on study findings. A good practice is for researchers to design their studies to be consistent with refutationist philosophy.[7-9] The central tenet of refutationist philosophy is to employ deductive logic in an effort to refute what one believes to be true. The proposed hypothesis generates logical predictions and then the researcher formally compares the collected data to those a priori predictions. Collected data that fail to refute the proposed hypothesis provide scientific evidence of its veracity. Scientists should make a genuine attempt to refute their proposed hypothesis rather than designing a study expected to generate data consistent with the hypothesis. Employing a refutationist approach, data collected from a single study could reject a hypothesis but results would not be sufficient to prove a particular scientific hypothesis. High-quality research provides evidence scientists can use to determine the best guess concerning the true state of nature in respect to a particular study.

Collected data that are an accurate representation of the true state of nature and sufficient to provide a genuine test of a proposed hypothesis provide quality scientific evidence. In general, this evidence relates to the estimation of an effect, or other estimator of interest, that is valid and preferably precise leading to statistical significance upon data analysis. An estimator is valid if multiple repetitions of the sample selection and data analysis processes yield a sampling distribution for the measure of interest that centres over the true population value.[10-12] Bias is a systematic error that causes the sampling distribution of the estimator to centre over a different value.[10, 13-15] A further definition of bias is a persistent error unrelated to imprecision or other sources of random error[16] in the data collection or analysis. Both random and systematic error influence study results (Figure 1) but random error can be visualised using confidence intervals whereas the effect of bias is often difficult to quantify. Data collected from a biased but precise measuring system will have the largest negative influence on study conclusions because the high precision might contribute to statistically significant findings. Increasing the sample size or performing multiple replications per sampling unit (within the context of the same study) can reduce the influence of random error on the measure of interest; however, these procedures will not reduce systematic error and the resultant bias. The validity, or relative lack of bias, of a study is seldom known with certainty since each study is only performed a single time and the true state of nature is typically unknown. Researchers must therefore apply logic to the evaluation of the study design and data collection procedures to conjecture whether or not important systematic errors were present within the study.

The three general categories of systematic errors that can affect studies of free-living populations are selection, information, and confounding bias (Table 1),[17, 18] which other authors have discussed in more detail elsewhere.[19-23] These classifications generally follow the three phases of observational studies: (a) the selection of subjects for study, (b) the collection of information from selected subjects, and (c) the analysis of collected data. The purpose of this article on bias was to introduce the required knowledge for researchers to evaluate the quality of evidence provided by published manuscripts within the *Equine Veterinary Journal*. This article discusses random error and each of the three major bias categories using examples of published manuscripts within the journal.

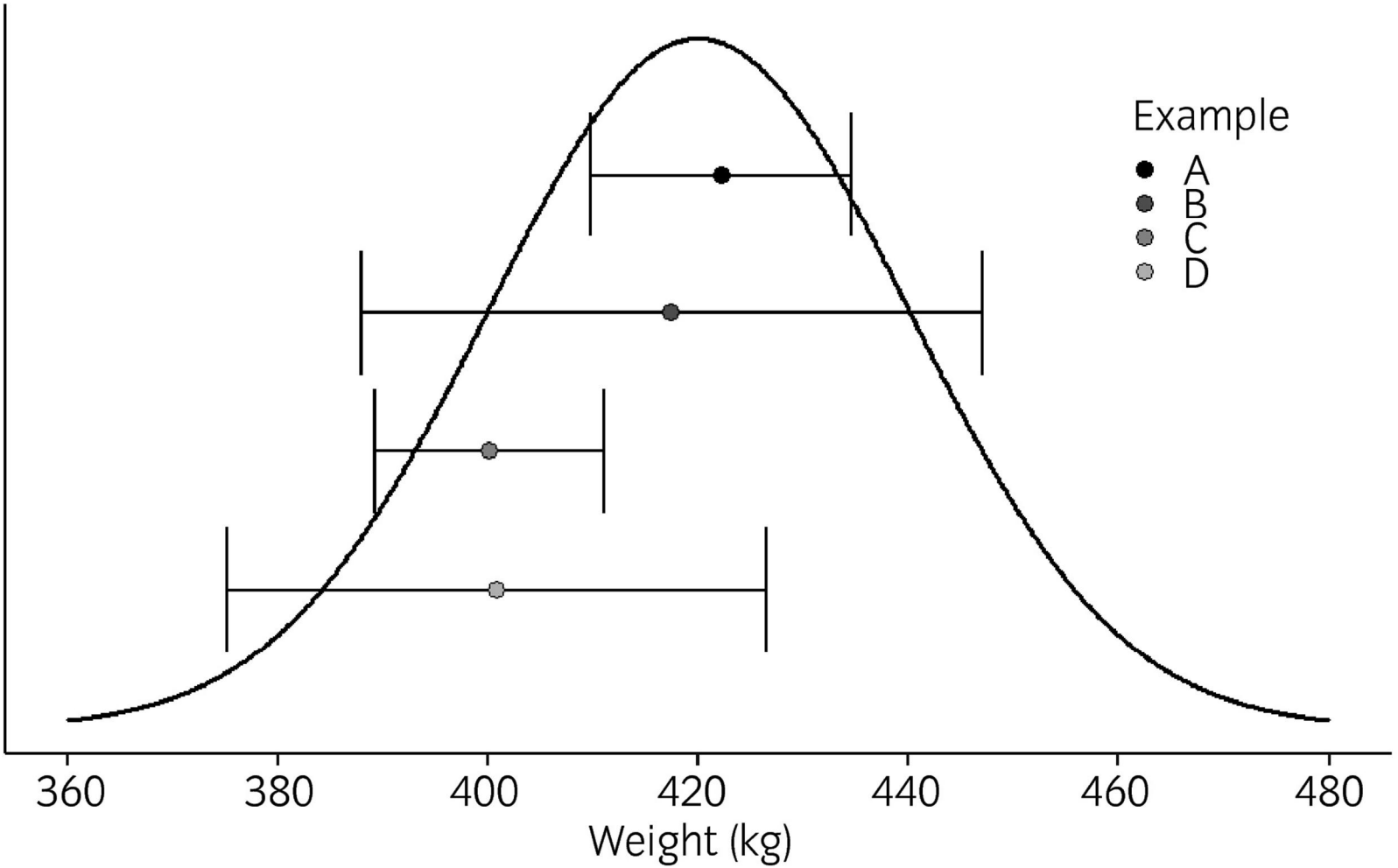

Figure 1. The mean and 95% confidence interval for the bodyweight of 10 randomly selected Nooitgedacht pony mares from a population with a mean of 420 kg and a standard deviation of 20 kg. Results are presented for a sample without systematic error (valid weight estimation) and without additional imprecision or random error (Example A), a valid weight measurement with additional imprecision (Example B), a biased weight measurement without additional imprecision (Example C), and a biased weight measurement with additional imprecision. Note that all confidence intervals contain the true population mean except for the biased but precise situation (Example C)

Table 1. Basic features of three general types of systematic errors that cause bias in the study of free-living populations

Bias category	Description	Occurrence	Mitigation
Selection	Animals included in the study are systematically different than the animals for which findings are to be generalised	Error in the methods employed to select study participants	Random selection of the target population
Information	Data collected from sampled animals are not an accurate reflection of the true state of nature	Error in the methods employed to collect data	Standardise data collection, collect objective rather than subjective data, train data collectors, employ blinding/masking
Confounding	The estimation of an effect does not represent reality due to the influence of an uncontrolled variable	Error in the data analysis by not accounting for the effect of all important variables	Collect appropriate data for all known variables that influence the study outcome and perform the appropriate statistical adjustments

2 RANDOM ERROR

2.1 General description

The two types of the errors than can occur within a study are random and systematic errors (see Figure 1). The presence of random errors causes imprecision in study estimates while systematic errors cause bias. Random errors typically develop from inherent variability of a measuring instrument but can also occur from observer variability and natural biological fluctuations. Variability of a measurement instrument could be due to temperature, humidity, and other external influences. The influence of measurement error on study findings is not always straightforward[24]; researchers should quantify the random error and resulting imprecision by calculating and presenting the standard deviations for the data and confidence intervals or standard errors for the point estimates of the effects. Studies presenting data from non-normal distributions should report the median with the absolute or interquartile range. Random error and imprecision of estimates have an important influence on statistical testing with low precision reducing the likelihood of identifying group variations as statistically different within commonly applied frequentist statistical approaches.

Precision and accuracy are related concepts but measuring precision is straightforward while accuracy is often unobservable except within the context of validation studies. An instrument with high precision will yield data that are repeatable with successive measurements of the same object being close to one another. In contrast, accuracy of an instrument is the relative closeness of the measured value to the true value. Accuracy of an individual measurement depends upon the precision of the instrument in addition to the lack of systematic errors (validity of the instrument). Systematic errors are frequently invisible whereas precision is easy to measure within the context of most studies. This dichotomy often leads to the subconscious belief that measurements with high precision are also accurate. Measurement systems with extensive random but no systematic error will still yield valid population estimates even if each individual measurement is not accurate due to imprecision. The primary influence of random error on study findings is that estimated population values will be imprecise (wide confidence intervals) and true differences between groups might not be statistically significant due to high measurement variability.

2.2 Mitigation procedures

The reduction of random errors in a study is relatively intuitive when viewed through the desire of identifying statistically significant findings. Increasing the sample size of the study will improve the precision of sample means (smaller standard errors of the mean and shorter confidence intervals) since most random measurement errors are normally distributed. Retaining the same sample size but increasing the number of sampling time points could be a feasible alternative depending upon the objective of an individual study. These are the primary methods to employ if the excessive variability is due to subject-specific factors such as natural biological variability. The use of repeated measurements on subjects and then averaging values for statistical analysis is a common method employed to account for random error inherent in the data collection instrument. Although, researchers could also analyse repeated measures data using appropriate statistical models including repeated measures ANOVA or mixed-effect statistical models. Appropriate training of observers and standardisation of measurement protocols are effective at reducing random errors that develop due to observer variability.

All studies will have random error and researchers should identify the important sources of random error during the design phase and implement appropriate mitigation procedures to reduce their impact on the results. The sample size for the primary objective of the study should be determined scientifically prior to initiating the study.[25] If feasible, investigators should assess the precision of measurement instruments and improve their performance prior to data collection. Authors should present point estimates with standard errors or confidence intervals within published manuscripts to provide readers with an appropriate description of data precision.

2.3 Study example

A manuscript by Gratwick *et al*. (2017) can be used to discuss issues related to random error and imprecision.[26] The objective of this study was "to compare the effects of a 4% modified fluid gelatin (MFG) with a 6% (130/0.4) hydroxyethyl starch (HES) on haemodilution, colloid osmotic pressure (COP), haemostasis and renal parameters in healthy ponies." Researchers selected six healthy Nooitgedacht pony mares from a university teaching herd for inclusion in a randomised crossover design.

The study evaluated two doses of modified fluid gelatin (10 ml/kg bodyweight and 20 ml/kg bodyweight) with a single dose of HES (10 ml/kg bodyweight). Investigators measured total serum protein (TSP), COP, haematocrit, and platelet counts at 1 hour prior to treatment (baseline), immediately post-treatment, and at 1, 2, 3, 6, 12 and 24 hours after treatment. Researchers evaluated coagulation parameters at baseline in addition to 1, 6, 12 and 24 hours post-treatment. The effects of treatment on kidney functioning was also investigated with specimens collected prior to the administration of each treatment, 24 hours after treatment, and 1 week after the final treatment administration. Presented descriptive statistics included the median and range due to the small sample size of six horses and apparent violation of the normality assumption for some of the measured variables. Researchers also transformed collected data prior to statistical analysis.

Mean haematocrit and TSP were different among treatment groups when evaluated over all time points but platelet counts were not different by treatment. Haemodilution should cause decreases in all three of these measurements so a difference with platelet values might be unexpected. Higher biological variation in platelet numbers within individuals (for example, due to platelet aggregation) might be the cause of this finding; however, these results suggest that the measurement of haematocrit and TSP have higher precision (greater repeatability) compared to methods used to estimate platelet numbers. The ranges for platelet counts (presented within the manuscript's Table 1) also suggest lower precision in these values. Researchers could have improved the precision of platelet estimation by performing the test multiple times on the same samples and recording the average count for statistical analysis. A modification of data collection would be important to consider for variables that are of prime importance to a given study.

The sample size of the current study was quite small and this might have been a reason for few significant differences between treatment groups when comparing data with less frequent sampling schedules. The study reported urinary function indices for only seven time points and the 1-week post treatment measures were the same sampling time as the baseline value prior to the subsequent treatment. The baseline values for enrolled horses varied descriptively among the three treatment groups despite the fact that these were the same horses enrolled in a crossover design. This finding suggests important biological variation within individual

horses (possible contributions include residual treatment effects) and for this reason, a larger sample size of horses would have been necessary to identify statistically significant differences due to this apparent source of random error. However, no horses developed clinical illness and all values remained within reference ranges suggesting that the identification of treatment differences for urinary function indices was not important within the context of this study.

3 BIAS CATEGORIES

3.1 Selection bias

3.1.1 General description

Selection bias[27] develops at the beginning of a study with the enrolment of animals and has been defined as "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".[14] Therefore, selection bias is present when the animals included in the study are systematically different from the population for which the results are desired to represent. Researchers should employ random sampling to obtain a representative study population but most studies will contain some degree of selection bias in absence of exhaustive sampling of the target population and this bias can develop due to a variety of mechanisms (Table 2).

Researchers must evaluate the potential impact of selection bias on the validity of presented results in conjunction with the study objective. For example, it is very likely that nonprobability sampling (ie not random including convenience, purposive and volunteer sampling) of a population of horses will cause meaningful bias in the estimation of prevalence in descriptive studies (Figure 2). If the objective of a particular study was to estimate some quality within a population of animals whether it is disease prevalence, presence of a potential risk factor for disease, or some other descriptive aspect then nonrandom sampling methods can be problematic. However, a cross-sectional investigation of the association between risk factors and disease might still yield valid risk factor results even if sampling methods bias the risk factor prevalence. Unacceptable levels of bias would be present in such a study if the effects of the risk factors on the disease within the sampled horses were different from the desired target population.

As an example, assume that a researcher is designing a study to investigate the effectiveness of administering mineral oil on the resolution of impaction colic. Furthermore, assume that the researcher wishes to extrapolate findings to horses with colic presenting to primary care practitioners. If the researcher selects horses presenting to a tertiary care facility for enrolment then the use of this study population could cause an unacceptable level of selection bias. This bias would occur because horses that present to tertiary care facilities are likely to have a different response to therapy than the colic cases handled by primary care veterinarians. Referral horses might be the more severe forms of disease that did not respond to previous therapy. A number of other factors are also likely to be different for horses presenting to a referral centre and these include time since onset, disease comorbidities, breed, and monetary value. Some of these factors could influence the effectiveness of therapy leading to additional selection bias. The use of nonprobability sampling methods is extremely common for observational studies and researchers must assess whether or not an unacceptable level of selection bias might be present on a case-by-case basis.

Table 2. Some specific types of selection bias that can influence the study of free-living populations

Bias name	Description	Influence	Design mitigation[a]
Nonresponse	Owners that fail to respond to the request for participation are systematically different than those that do respond	Respondents will not be representative of the entire target population of interest	Design simple, quick data collection instruments (eg questionnaires). Employ multiple follow-up attempts and consider adding an incentive to improve response proportions.
Volunteer	The animals of owners that volunteer for a study are systematically different than the target population of animals	Management of participating animals might be better in general and animals might be in better overall health	Design the study objective with this consideration and employ strict inclusion and exclusion criteria. Consider adding incentives to participation.
Loss-to-follow-up	The animals that cannot be followed for the entire duration of the study are different in important characteristics than the animals completing the study	Management of animals that remain might be better and animals might be in better overall health. Losses could be due to illness or other competing interests	Education of owners concerning the purpose of the study. Strict inclusion criteria concerning animal and owner characteristics. Incentives to reduce losses during the study
Disease spectrum	Animals selected for study do not have a representative range of disease severity	Response to treatment will be over- or under-estimated depending on the types of cases that are enrolled	Careful selection of the study population and appropriate inclusion and exclusion criteria
Berksonian	Animals enrolled in case-control studies are more likely to have both the exposure and disease of interest. This is most common when cases are selected from tertiary care facilities	Enrolled cases might be more likely to have certain co-morbidities. For example, valuable horses with comorbidities are more likely to present to referral practices	Careful selection of the study population and inclusion criteria for cases and controls

[a] Researchers could apply post hoc statistical adjustments to mitigate the negative effects of selection bias in some situations when appropriate validation information is available

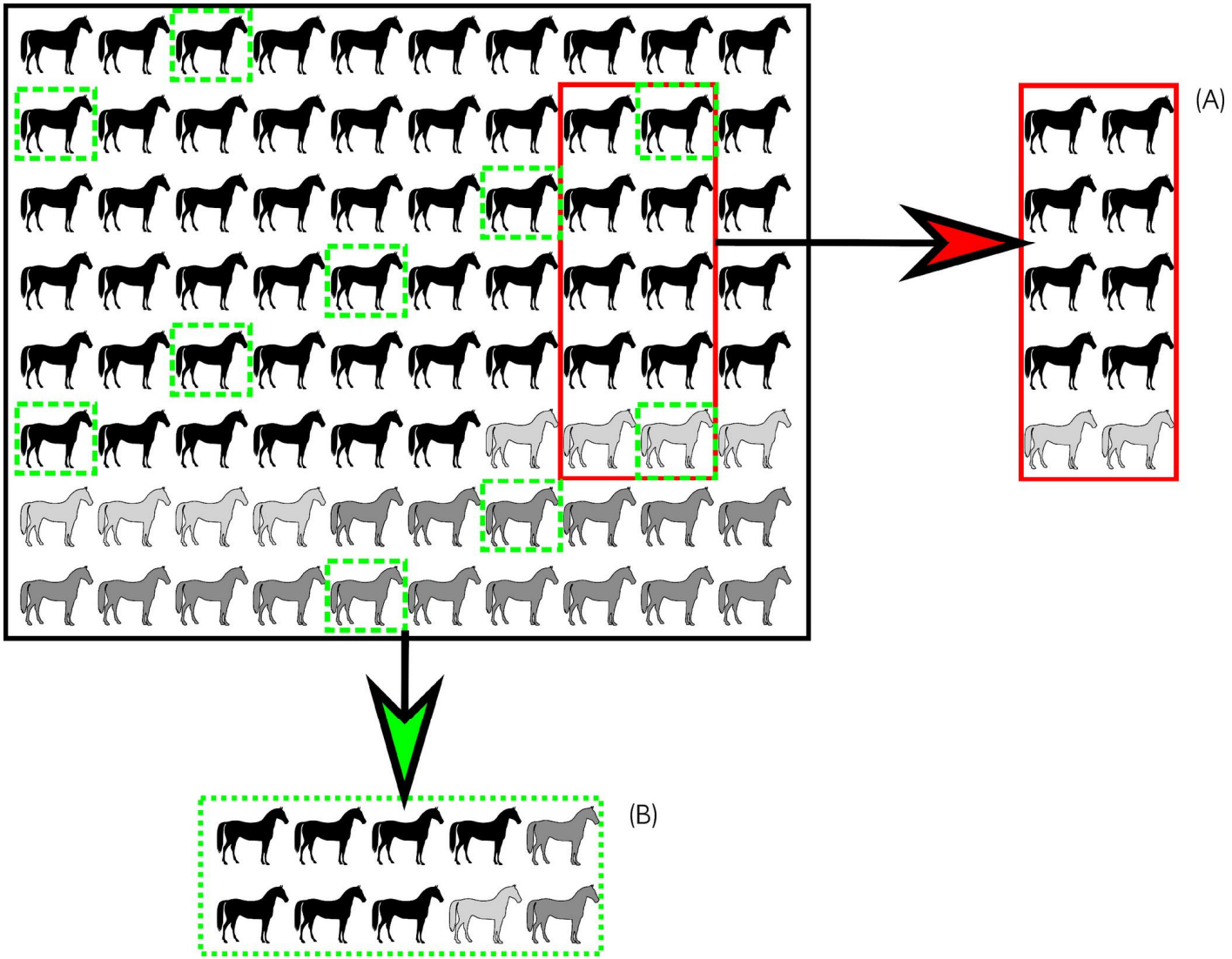

Figure 2. Nonrandom convenience sampling of horses causing selection bias (A) in respect to the prevalence of horse colours compared to a valid sample obtained via systematic random sampling (B)

3.1.2 Mitigation procedures

Probability sampling of the target population of horses for enrolment into the study population is the only method that is guaranteed to prevent selection bias. Random sampling is seldom a feasible option and researchers must therefore purposely select a population of horses that is representative of the desired target population in respect to the primary study objective. After completion of the study, researchers will have difficulty making the appropriate corrections for selection bias.[28] It is therefore imperative that researchers give careful thought to the influence of this bias during the design and subsequent animal enrolment phases of the study.

3.1.3 Study example

A manuscript by Viljoen *et al*. (2014) can be used to discuss issues related to selection bias.[29] The objective of this study was "to evaluate the haemostatic and oncotic effects of tetrastarch (130/0.4) administered at 10, 20 and 40 ml/kg bwt in healthy horses." Investigators studied six healthy Nooitgedacht pony mares from a university teaching herd. The Nooitgedachter was developed in South Africa as a riding horse and is considered a rare breed.[30, 31]

The six Nooitgedacht mares were enrolled in a randomised crossover design to investigate the effects of different dosages of hydroxyethyl starch. Selection bias is not considered an important concern of randomised clinical studies when randomisation has been performed appropriately.[32] However, selection bias might pose a problem when considering the extrapolation of findings to the target population of interest. There are two issues to consider relative to the study population in this instance: (a) Nooitgedacht is a rare breed and (b) healthy horses are not the typical recipient of hydroxyethyl starch therapy. The objective states healthy horses in general; however, researchers did not select a random sample of healthy horses and a representative sample of all healthy horses would not exclusively be Nooitgedacht ponies. Selection bias is therefore a potential concern with the findings and breed variability in the haemostatic and oncotic effects of hydroxyethyl starch administration would cause a selection bias. However, healthy horses of all breeds likely have similar physiological responses and therefore the first concern related to the selection of horses was unlikely to cause meaningful selection bias. Additionally, the study of healthy horses could not be a source of selection bias per se because of the stated objective. Although, the reader of the manuscript that wishes to incorporate presented findings into their clinical practice must decide if data collected from healthy horses can reasonably predict what would happen in clinically ill horses. It is standard practice to evaluate therapies in clinically healthy animals prior to performing studies of patients and therefore this potential limitation does not reduce the usefulness of the findings.

3.2 Information bias

3.2.1 General description

Information bias[33] develops when data collected from the sampled animals deviate from the true information and there are a number of classifications and sources of information bias that can affect studies (Table 3). Information bias develops from mechanisms independent of imprecision in data collection methods, which will add random but not systematic error. Imprecision in the measurement of a factor in absence of systematic error will cause misclassification that will bias results towards the null (smaller effect measures) when the

Table 3. Some specific types of information bias that can affect studies of free-living populations

Bias name	Description	Influence	Design mitigation[a]
Nondifferential misclassification	The amount of misclassification of either exposure or disease status does not vary based on comparison groups	In a 2 × 2 table situation, the measure of association will be biased towards the null (towards "no effect")	Standardised data collection procedures, independent data extraction by two researchers, pretest data collection instruments
Differential misclassification	The amount of misclassification is different among the comparison groups	The direction of the bias cannot be predicted	Same as for nondifferential misclassification with the additional requirement of employing blinding or masking of data collectors
Recall	Owners of animals do not accurately remember information from the past	The direction of the bias cannot be predicted but it is more damaging to study validity when the magnitude of the bias (or direction) varies by comparison groups	Pretest questionnaire and consider removing questions based on long time delays. Triangulate data using multiple data sources or questions.
Detection	A type of differential misclassification of disease status when the probability of disease detection is influenced by the presence of another factor	Results will be biased away from the null when the disease is more likely to be detected in populations with the characteristic under study	Standardised data collection, training of observers, and the use of blinding to prevent differential detection
Self-reporting	Respondents to questionnaires or interviews will provide socially acceptable answers or the answers they feel are expected by the researchers	Prevalence will be over- or under-estimated depending upon the socially acceptable response. Typically causes nondifferential misclassification and bias towards the null	Careful questionnaire design with pretesting to detect leading questions or questions with language that suggest one type of response is more acceptable than another
Interviewer	The person performing interviews elicits biased data through the manner with which questions are asked	Prevalence will be over- or under-estimated depending upon the beliefs of the interviewer. Might cause differential or nondifferential misclassification	Training of interviewers and the use of standardised scripts for asking questions. Blind interviewer to the exposure status of the group being interviewed to reduce the likelihood of differential errors in data collection
Observer	Collected data are systematically different from the truth due to conscious or subconscious predispositions of the observer	Prevalence will be over- or under-estimated depending upon the predisposition of the interviewer. Might cause differential or nondifferential misclassification	Training of observers and the use of standardised data collection methods. Blind observers to the exposure status to reduce the likelihood of differential errors
Lead-time	The time from disease onset until detection is different in the study compared to what normally occurs. Typically used to describe differential errors across study groups	Survival times might be over- or under-estimated depending upon the enrolment of animals. In some instances could be described as a selection bias or alternatively cause confounding depending upon study objective	Employ standardised testing protocols that are consistent with previous research. Consider blinding to ensure that the detection of disease is not differential among comparison groups

[a] Researchers could apply post hoc statistical adjustments to mitigate the negative effects of information bias in some situations when appropriate validation information is available.

factor is dichotomised[34] but results are unpredictable when data are categorised into more levels.[16, 24]

Researchers must evaluate the potential impact of information bias in terms of the stated study objective. For example, a descriptive study estimating the prevalence of disease might have limited scientific value if substantial information bias were present. However, the impact might be less severe for analytical studies that aim to compare two or more groups as long as the errors in information were the same among all groups to be compared (nondifferential). Misclassification[35-39] of predictors or outcomes is a common consequence of information bias. Differential misclassification of risk factors or disease status is an error where the magnitude varies by study groups and the overall effect of the bias is difficult to predict. However, nondifferential misclassification of dichotomised data will bias results towards the null. In such instances, statistically significant findings reported in the study would not be a consequence of the information bias and actually would be an underestimation of the true population effect. Therefore, assuming that researchers were investigating group differences, then nondifferential misclassification would not invalidate conclusions nor should readers disregard the findings as unacceptably flawed.

As an example, assume that a researcher is investigating risk factors for the development of Tyzzer's disease in foals.[40] On a particular Thoroughbred breeding farm, there were nine diagnosed cases of Tyzzer's disease during a 3-year period. Investigators randomly selected 54 nonaffected foals present on the farm to provide a 1:6 ratio of cases to controls. Only one affected foal was born to a resident mare and univariate statistical analysis estimated that the odds of having a Tyzzer's affected foal to be four times greater for visiting compared to resident mares (odds ratio [OR] = 4). However, had there been errors in extracting data from the records and the resident status of mares was incorrectly recorded for 12% of the foals then this nondifferential misclassification would have caused the OR to reduce to 2 for visiting mares or affected foals being twice as likely to be from visiting dams (Figure 3). This attenuation of the measure of association (bias towards the null value of OR = 1) is the hallmark of nondifferential misclassification (within a 2×2 table situation). This particular misclassification is analogous to a random measurement error where the sensitivity and specificity for correct classification was equal (here, sensitivity = specificity = 88%).

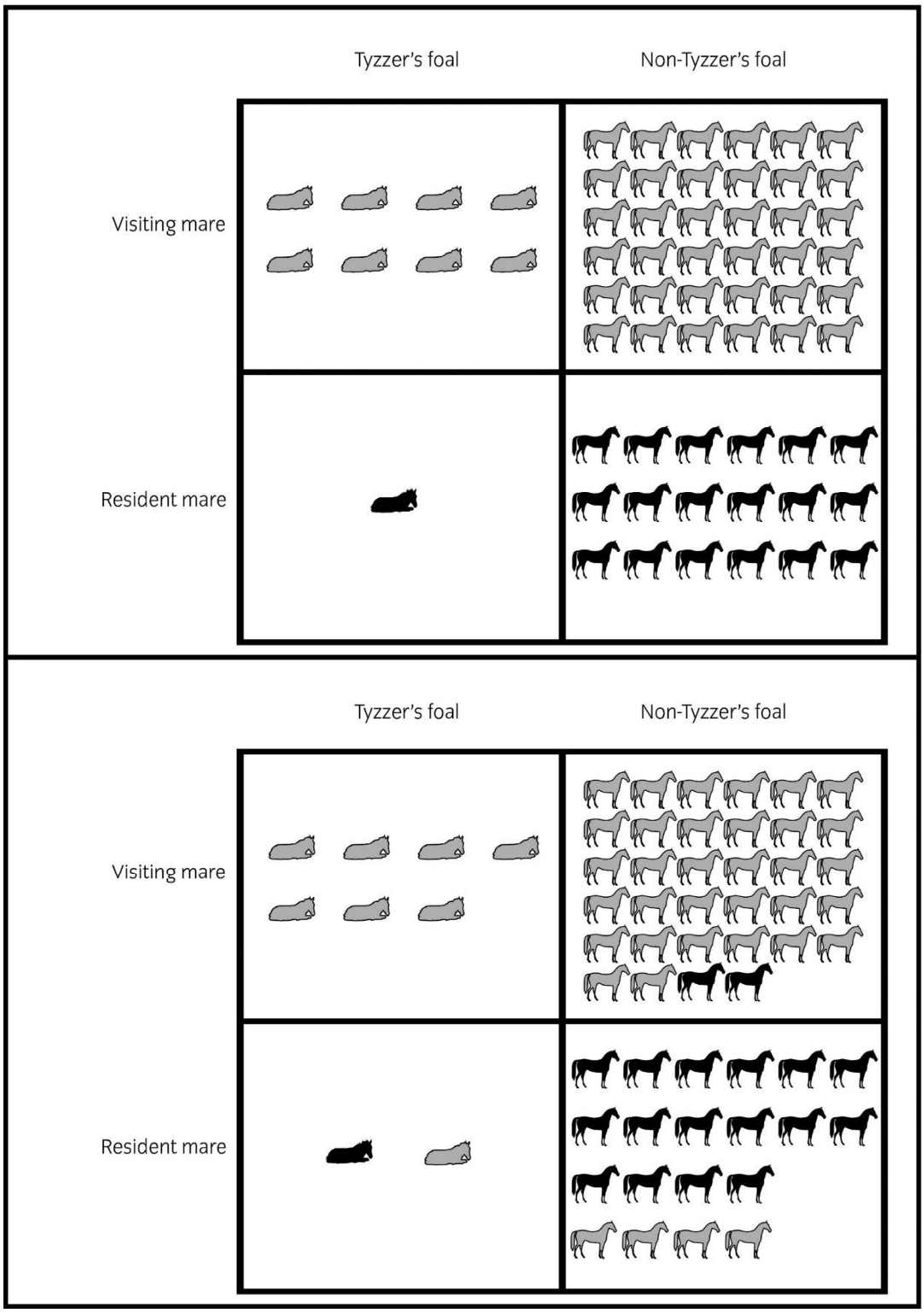

Figure 3. The 2x2 tabular analysis comparing the relationship of mare residence status on the diagnosis of Tyzzer's disease in her foal assuming the published findings were the true state of nature (top panel) compared to 12% nondifferential misclassification of exposure status (bottom panel). The odds for Tyzzer's foals to have visiting mares were four times the odds of Tyzzer's foals to have nonvisiting mares (95% confidence interval; 0.46-34.5; top panel). Due to misclassification, the odds for Tyzzer's foals to have visiting mares were two times the odds Tyzzer's foals to have nonvisiting mares (95% confidence interval; 0.39-10.9; bottom panel)

3.2.2 Mitigation procedures

The majority of studies making observations on free-living populations will have some degree of information bias. Systematic errors that are differential among comparison groups are the most damaging and careful attention to the study design is required to limit their influence. The collection of objective data using validated instruments by trained observers is essential to reduce information bias. Researchers should pretest questionnaires and train administrators to ensure standardised data collection. The use of blinding or masking of group assignment is necessary when data to be collected include subjective observations. It is essential that data are collected in a consistent manner across all study groups and corresponding to a consistent timeframe (eg time from diagnosis). The statistical analysis cannot mitigate the negative effects of information bias except in rare instances when validation data are available to adjust for misclassification or measurement error.

3.2.3 Study example

A manuscript by McConnell *et al*. (2013) can be used to discuss some issues related to information bias.[41] The objective of this study was "to determine whether near infrared spectroscopy (NIRS) can identify trends in regional cerebral oxygen saturation (rSO_2) in horses and whether there is a correlation between rSO_2 and venous oxygen tensions." The study design included six healthy Nooitgedacht mares from a university teaching herd. Researchers randomised horses into two groups for the ordering of hyper- and hypocapnia conditions under general anaesthesia. Investigators collected data corresponding to eight, 10-minute time periods during the course of the study.

Within the six study mares, researchers measured rSO_2, arterial partial pressure of oxygen (PaO_2), venous partial pressure of oxygen (PvO_2), arterial partial pressure of carbon dioxide ($PaCO_2$), and venous partial pressure of carbon dioxide ($PvCO_2$) during standing (recording period [RP] 1), standing sedation (RP 2), general anaesthesia with normocapnia (RP 3), general anaesthesia with hypercapnia (RP 4), general anaesthesia with hypocapnia (RP 5), completion of general anaesthesia (RP 6), 5 minutes post-recovery (RP 7) and 2 hours post-recovery (RP 8). Investigators fitted the rSO_2 monitor over the dorsal sagittal sinus of each horse for monitoring throughout the study. Researchers measured blood gases three times at 1 min, 5 min and 10 min of each RP. Blood collection and blood gas measurements followed standard procedures and were therefore unlikely to be a source of meaningful information bias. However, information bias might have influenced rSO_2 measurements since there is no available gold standard measurement. Previous studies have successfully incorporated cerebral rSO_2 monitoring in human patients[42] despite the lack of a gold standard; however, the usefulness of monitoring equine patients had not been determined.

There are obvious anatomical differences between horses and people and it is unclear if the different probe positioning and thickness of skin and bone influenced calculated rSO_2 measures. The concentration of both oxygenated and deoxygenated haemoglobin is calculated using light optical spectroscopy for determination of rSO_2. The attenuation of light varies with tissues of different thickness and density and it is uncertain if the algorithm used to calculate rSO_2 is robust to such differences between horses and people. Also, the dorsal sagittal sinus receives primarily venous blood, which might be a different proportion than what the algorithm was designed to measure from the frontal cortex in human patients. The high variability among data collected from different horses (presented in the article's Table 2) suggests that individual anatomical variations might have influenced rSO_2 measurements.

The lack of significant differences among procedures during general anaesthesia and the moderate correlations with blood gas parameters highlights potential validity concerns related to the rSO_2 data.

There are currently no guidelines on how to evaluate rSO_2 values concerning cerebral oxygenation desaturation and the requirement for intervention in anaesthetised horses. This study was an exploratory evaluation that reported rSO_2 variation over time in healthy horses suggesting the potential as an anaesthetic monitoring tool. The presence of information bias might compromise the validity of reported findings. However, any information bias in study measurements would have applied equally to all recording periods and therefore the observed significant differences should be robust findings. Information bias that differs among comparison groups is the most damaging type and this was not present in the current study. Anatomical variations among horses might have caused increased random errors in addition to potential systematic errors and this could have further contributed to the few significant differences between experimental time points. A larger sample size of horses might have provided sufficient data to identify a greater number of significant differences.

3.3 Confounding

3.3.1 General description

Confounding bias[43] manifests during the analytical phase of the study after the completion of data collection. Confounding is the mixing of effects between three or more variables when the researcher desires to estimate a valid association between a single predictor and an outcome. Confounding bias is present when the association measured between the predictor and the outcome deviates from the true value due to the influence of one or more other variables. All analytical observational studies likely contain some degree of confounding and appropriate adjustment during the statistical analysis is required to reduce the resultant bias.

Researchers must evaluate the potential impact of confounding within a study based on their biological understanding of the factors involved in the outcome (eg disease) under investigation. Confounding can affect studies that evaluate risk factors for disease or make predictions related to another health outcome. Confounding cannot affect the results of descriptive studies that do not measure the association between two variables. Three 'rules of 3' are important considerations when evaluating the potential for confounding bias in a study. There must be at least three variables under consideration: (a) a disease (outcome) of interest, (b) a potential risk factor (exposure/predictor of primary importance) of the disease of interest, and (c) one or more confounding variable that might influence the measured association between the potential risk factor and disease. The second 'rule of 3' describes the criteria for the development of confounding: (a) the potential confounder must be *causally related* to the outcome under study, (b) the potential confounder must be associated with the exposure of interest *within the dataset being analysed*, and (3) the potential confounder must not be on the causal pathway between the exposure and outcome of interest. The third 'rule of 3' corresponds to confounding being potentially from (a) an unknown, (b) known but unmeasured, or (c) measured but statistically unadjusted variable within the analysis. It is also possible for residual confounding to be present within results in which statistical adjustment was insufficient to eliminate all bias.

The adjustment for confounding when the three criteria for its development (second 'rule of 3') have not been satisfied can actually increase, rather than decrease bias. For example, a

collider is an intermediary variable independently caused by both the exposure and outcome of interest.[44] A collider variable will likely be associated with the exposure of interest due to this causal association and will not be on the causal pathway leading from the exposure to the outcome of interest. However, it will also not be a causal risk factor for the outcome of interest (causal relationship is the reverse) and therefore the control of this variable in the statistical analysis can introduce bias rather than correct for it. A thorough understanding of the underlying biology of the outcome of interest is a necessary requirement to evaluate the potential for confounding bias. Investigators can describe the underlying biology using directed acyclic graphs, which will subsequently facilitate the appropriate adjustment of confounding.[44-48]

Simpson's paradox[49] is the confounding of a relationship that is so extreme that the measured effect is opposite to the true state of nature. For example, the use of an inactivated vaccine for the prevention of equine strangles might appear to be a cause of strangles if veterinarians preferentially administer the vaccine to high-risk, exposed foals. This would be an example of "confounding by indication"[50] where the status of the enrolled animal influences the therapy that is administered. Confounding occurs when the indication to provide the therapy is a risk factor for the outcome under investigation. Similarly, it is also possible that clinicians consider the severity of illness when deciding on the appropriate treatment regimen. This will also create the potential for confounding to be present when comparing the effectiveness of different treatment options. Confounding is a serious concern of clinical studies when the exposure of interest (eg a treatment) is not randomly allocated to study subjects.

As an example, consider the previously mentioned study investigating risk factors for Tyzzer's disease in foals present on a Thoroughbred breeding farm.[40] The foals of visiting mares had a four times higher odds (OR = 4) of being diagnosed with Tyzzer's disease compared to foals born to resident mares based on the crude unadjusted statistical analysis. However, susceptibility for Tyzzer's disease varies by age[51] and this is a potential confounder of the measured association (Figure 4). Within the study population, the average age of foals when visiting mares arrived at the breeding farm was 13 days (range, 0-77 days) compared to a mean of 0 days (range, 0-13) for mares that were resident on the farm. Foals are susceptible to clinical disease up until 6 weeks of age with most cases reported within 2-4 week old foals.[52] Foal age at arrival is a potential confounder of the estimated relationship between dam residence status and Tyzzer's disease in her foal. Confounding could be present because (a) foal age is a predisposing risk factor for Tyzzer's disease, (b) foal age varied by mare residence status within the study population, and (c) foal age is not on the causal pathway between mare residence status and development of Tyzzer's disease. It is therefore possible that the reported crude OR between mare residence status and Tyzzer's disease is biased because the authors did not account for foal age during the statistical analysis. However within this study population, the actual amount of bias induced by the effect of age is likely negligible due to the fact that only a single foal born to a resident mare developed Tyzzer's disease. Furthermore, foals of visiting mares tended to be older so the OR is likely biased towards the null and smaller than the true population value.

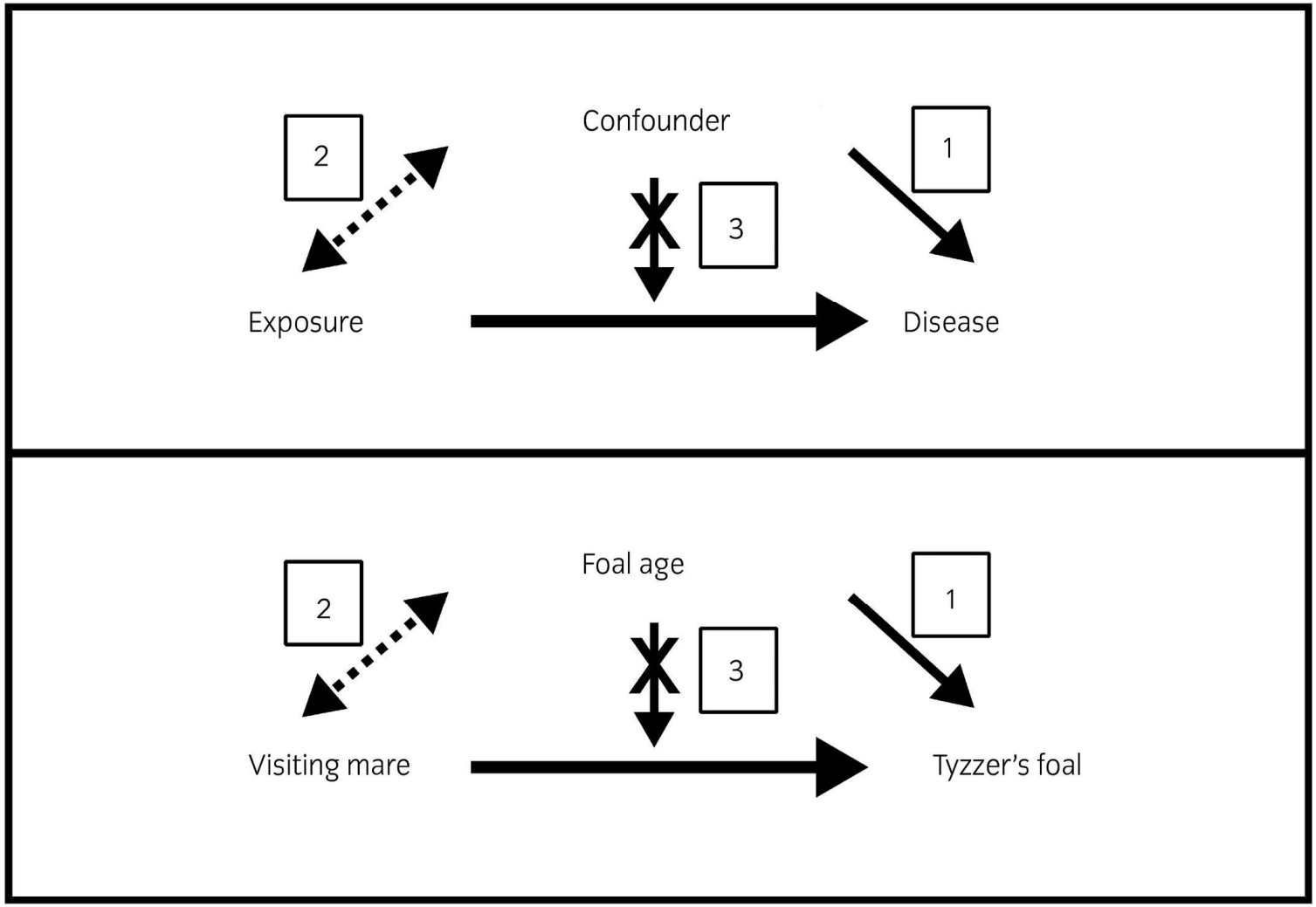

Figure 4. A general diagrammatic representation of the criteria for confounding (top panel) and a hypothetical example of the confounding effect of foal age on the diagnosis of Tyzzer's disease in foals born to mares visiting a Thoroughbred breeding farm (bottom panel). Bias will be possible if the potential confounder is a causal risk factor for the outcome under study (1), there is a mathematical relationship within the current data between the confounder and the exposure under study (2), and the confounder is not an intermediary variable along the causal pathway from exposure to outcome (3)

3.3.2 Mitigation procedures

Researchers should assume that all studies of free-living animals contain some degree of confounding. Randomisation of treatment allocations within a randomised controlled study will prevent confounding for known and unknown factors because it breaks the link between the potential confounder and the primary exposure of interest (criteria 2 depicted in Figure 4). However, confounding will remain a concern in small studies employing randomisation since the procedure is unlikely to create equal distributions for all potential confounders. Confounding might therefore still be a possibility within the typical randomised control study performed in equine health.

Matching is another design feature used to address confounding in epidemiological studies. However, matching controls to cases within a case-control framework will create confounding if the appropriate statistical analysis is not performed to adjust for this matching (eg matched analysis). Case-control studies employ matching because it improves the efficiency of confounding control for the matched factors during the statistical analysis. Matching within a cohort study design will prevent confounding by the matched factors and no special statistical analysis will be required. The different effect of matching between case-control and cohort designs is that matching in a cohort design makes the animals within each exposure category similar (removes criteria 2 depicted in Figure 4). In contrast, matching cases to controls in a case-control design will actually induce a mathematical relation between the matching factor and the exposure of interest (assuming both matching factor and exposure of interest are causally associated with the disease under study). Analysts must therefore correct for the matching during the statistical analysis. Researchers should only match for variables that are established risk factors for the outcome under study and when the researchers have no desire to estimate the effects of the matching variable, which will not be possible when performing the matched analysis.

Restriction is another effective approach for the prevention of confounding bias. For example, if the majority of cases of Tyzzer's disease occurs in foals 2-4 weeks of age then the study population could be restricted to only younger foals. This restriction will prevent confounding by foal age and subsequent statistical adjustments will not be required. Restriction is most effective when high-quality information is available concerning the potential confounder and when it is feasible to restrict the study population in this manner. In practice, restriction typically reduces confounding but researchers should still implement appropriate statistical methods to remove any residual confounding bias that might be present within collected data.

The most common approach is to control for confounding during the statistical analysis of the data. Confounding can be controlled using stratified tabular analyses and the calculation of weighted averages of the stratum-specific measures of association (eg OR). Disease occurrence estimates can also be standardised[53] across populations for comparison purposes; however, this is an uncommon practice in veterinary medicine. The most common approach to adjust for confounding is the use of multivariable statistical models. Inclusion of both the exposure of interest and the potential confounder in the same statistical model will effectively adjust the measure of effect for confounding. The multivariable model can include all potential confounders to yield the most valid effect estimate for the primary exposure of interest. However, the inclusion of variables that are not causal risk factors for the outcome under investigation might introduce rather than remove confounding. Researchers should carefully consider the biology related to the variables included in statistical models for this

reason. Although multiple potential confounders can be included, increasing the number of variables in statistical models will reduce the overall precision of results. The control of bias must always take precedence but analysts should consider removing potential confounders from statistical models when substantial confounding is not present within the data set under investigation.

Analysts should estimate the magnitude of confounding bias by calculating the percent change in the measure of association between the statistical model including the potential confounder (adjusted) and the model that does not include the confounder (unadjusted).

$$\text{Percent change} = [\,(\text{adjusted measure of association} - \text{unadjusted measure of association}) \,/\, \text{adjusted measure of association}$$

Meaningful confounding is present within the data if the absolute value of the percent change is ≥20%. Although in larger studies, a percent change ≥10% might be more appropriate. Analysts could remove the potential confounder from the statistical model if the bias does not reach this threshold in an effort to estimate more precise effect measures. However, researchers could also retain potential confounders in statistical models even when minimal confounding is present if adjustment for specific confounders is expected within the field of study. Researchers should estimate the amount of bias and not use statistical tests to inform decisions related to the retention of potential confounders within statistical models.

3.3.3 Study example

A manuscript by Joonè *et al*. (2017) can be used to discuss issues related to confounding.[54] The objectives of this study included an assessment of the contraceptive efficacy of native porcine zona pellucide (pZP) vs recombinant porcine zona pellucida (reZP) vaccines. The researchers obtained 21 Nooitgedacht pony mares from a university teaching herd and then stratified mares by age before randomly assigning seven to each of three treatment groups: (a) pZP vaccination, (b) reZP vaccination, and (c) adjuvant-only control. Investigators performed transrectal palpation and ultrasonography and collected whole blood for serum separation every 7 days during the study. Researchers observed mares for oestrus behaviour and an experienced reproduction specialist bred cycling mares by artificial insemination using fresh semen from a single stallion starting 5 weeks after the completion of the vaccination schedule. No mares in the pZP group subsequently became pregnant (0/7) compared to 57% (4/7) and 100% (7/7) in the reZP and placebo control groups, respectively. The pregnancy proportion was significantly different between the pZP and control groups but not any other pairwise comparison.

The study employed random allocation after stratifying by age and this procedure is the best method to prevent confounding bias. However, the small size of the study suggests that it would be theoretically possible for the unequal distribution of potential confounders among treatment groups. For example, the baseline characteristics of the study groups (presented in the manuscript's Table 1) indicate that mares in the control group tended to be younger with a heavier body weight relative to the other two groups. This observation might suggest that control mares would have a greater likelihood of becoming pregnant during the study irrespective of treatment administration. The lack of significant differences in potential confounders among groups does not preclude the possibility of confounding bias. In general, the descriptive variation in potential confounding variables suggests that the statistical analysis should investigate and potentially control for these variables. Practically, the 0% and 100% pregnancies between the pZP and placebo groups indicate that these particular

variables could not be the primary reason for the observed group differences. It is theoretically possible that a proportion of mares in the university teaching herd was infertile due to some undiagnosed reason and the unequal distribution of this variable is the true cause of the observed differences among groups. This type of situation is unlikely considering the intensive examinations of the study horses but the intermediate effect of reZP might be due to confounding by an unmeasured variable unevenly distributed by the randomisation procedures. Researchers might therefore consider repeating the study within a larger population of horses to confirm these findings.

4 CONCLUSIONS

The purpose of this article on bias was to introduce the required knowledge for readers of the journal to evaluate published manuscripts in respect to the presence of random and systematic errors. Researchers experienced at performing critical reviews will produce higher-quality research, as they will be aware of the common sources of study errors. The use of reporting guidelines can enhance the quality of manuscripts but consumers of research must still apply logic in the evaluation of published reports. There is no perfect study, all will have some degree of random and systematic errors and biased data with high precision are the most damaging to study inferences. Random errors should be quantified using measures of dispersion and confidence intervals. In contrast, systematic errors are difficult to quantify and evaluation of these errors must consider the primary study objective. The appropriate design of the study is the most important aspect to reduce the negative impacts of selection and information bias. Good study design practices can also reduce confounding bias, but in most instances, analysts perform these adjustments during the statistical analysis. A thorough understanding of both biology and statistical model building is necessary to control confounding appropriately. There are many biological and analytical considerations when designing studies of free-living populations and it is unrealistic to expect an individual study to be perfect. Authors are therefore encouraged to facilitate the critical review process by providing text related to the limitations and potential sources of bias within the discussion section of their manuscripts.

CONFLICT OF INTERESTS

No competing interests have been declared.

REFERENCES

1. Bornmann L, Mutz R. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *J Assoc Inf Sci Technol*. 2015; 66: 2215– 22.

2. Larsen P, Ins M. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*. 2010; 84: 575– 603.

3. Carrell DT, Simoni M. 'Easier ways to get a publication': the problem of low quality scientific publications. *Andrology*. 2018; 6: 1– 2.

4. Miller J. Where does the time go? An academic workload case study at an Australian University. *J High Educ Policy Manag*. 2019; 41: 633– 45.

5. Simera I, Moher D, Hoey J, Schulz KF, Altman DG. A catalogue of reporting guidelines for health research. *Eur J Clin Invest*. 2010; 40: 35– 53.

6. Wang X, Chen Y, Yang N, Deng W, Wang Q, Li N, et al. Methodology and reporting quality of reporting guidelines: systematic review. *BMC Med Res Methodol*. 2015; 15: 74.

7. Buck C. Popper's philosophy for epidemiologists. *Int J Epidemiol*. 1975; 4: 159– 68.

8. Popper KR. Science: problems, aims, responsibilities. *Fed Proc*. 1963; 22: 961– 72.

9. Weed DL. An epidemiological application of Popper's method. *J Epidemiol Community Health*. 1985; 39: 277– 85.

10. Brenner H. RE: "Does nondifferential misclassification of exposure always bias a true effect toward the null value?". *Am J Epidemiol*. 1991; 134: 438– 9.

11. Mertens TE. Estimating the effects of misclassification. *Lancet*. 1993; 342: 418– 21.

12. Szklo M, Nieto FJ. Epidemiology: beyond the basics. Gaithersburg, MD: Aspen; 2000. p. 125– 6.

13. Flegal KM, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. *Am J Epidemiol*. 1991; 134: 1233– 44.

14. Fosgate GT, Cohent ND. Epidemiological study design and the advancement of equine health. *Equine Vet J*. 2008; 40: 693– 700.

15. Rothman KJ. Epidemiology: an introduction. Oxford, New York: Oxford University Press; 2002. pp viii, 223.

16. Fosgate GT. Non-differential measurement error does not always bias diagnostic likelihood ratios towards the null. *Emerg Themes Epidemiol*. 2006; 3:7.

17. Ibrahim MA, Spitzer WO. The case control study: the problem and the prospect. *J Chronic Dis*. 1979; 32: 139– 44.

18. Schwartz S, Campbell UB, Gatto NM, Gordon K. Toward a clarification of the taxonomy of "bias" in epidemiology textbooks. *Epidemiology*. 2015; 26: 216– 22.

19. Christenfeld NJ, Sloan RP, Carroll D, Greenland S. Risk factors, confounding, and the illusion of statistical control. *Psychosom Med*. 2004; 66: 868– 75.

20. Greenland S, Morgenstern H. Confounding in health research. *Annu Rev Public Health*. 2001; 22: 189– 212.

21. Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet*. 2002; 359: 248– 52.

22. Miettinen OS, Cook EF. Confounding: essence and detection. *Am J Epidemiol*. 1981; 114: 593– 603.
23. Sackett DL. Bias in analytic research. *J Chronic Dis*. 1979; 32: 51– 63.

24. van Smeden M, Lash TL, Groenwold RHH. Reflection on modern methods: five myths about measurement error in epidemiological research. *Int J Epidemiol*. 2020; 49: 338– 47.

25. Fosgate GT. Practical sample size calculations for surveillance and diagnostic investigations. *J Vet Diagn Invest*. 2009; 21: 3– 14.

26. Gratwick Z, Viljoen A, Page PC, Goddard A, Fosgate GT, Lyle CH. A comparison of the effects of a 4% modified fluid gelatin and a 6% hydroxyethyl starch on haemodilution, colloid osmotic pressure, haemostasis and renal parameters in healthy ponies. *Equine Vet J*. 2017; 49: 363– 8.

27. Kleinbaum DG, Morgenstern H, Kupper LL. Selection bias in epidemiologic studies. *Am J Epidemiol*. 1981; 113: 452– 63.

28. Hanley JA. Correction of selection bias in survey data: is the statistical cure worse than the bias? *Am J Epidemiol*. 2017; 185: 409– 11.

29. Viljoen A, Page PC, Fosgate GT, Saulez MN. Coagulation, oncotic and haemodilutional effects of a third-generation hydroxyethyl starch (130/0.4) solution in horses. *Equine Vet J*. 2014; 46: 739– 44.

30. Joubert DM, Bosman WM. The Nooitgedacht Pony. *S Afr J Sci*. 1971; 67: 8.

31. van der Merwe FJ, Martin J. Four Southern African horse breeds. *Anim Genet Resour Inf*. 2002; 32: 57– 72.

32. Kahan BC, Rehal S, Cro S. Risk of selection bias in randomised trials. *Trials*. 2015; 16: 405.

33. Kesmodel US. Information bias in epidemiological studies with a special focus on obstetrics and gynecology. *Acta Obstet Gynecol Scand*. 2018; 97: 417– 23.

34. Hofler M. The effect of misclassification on the estimation of association: a review. *Int J Methods Psychiatr Res*. 2005; 14: 92– 101.

35. Birkett NJ. Effect of nondifferential misclassification on estimates of odds ratios with multiple levels of exposure. *Am J Epidemiol*. 1992; 136: 356– 62.

36. Brenner H, Loomis D. Varied forms of bias due to nondifferential error in measuring exposure. *Epidemiology*. 1994; 5: 510– 7.

37. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol*. 1977; 105: 488– 95.

38. Gladen B, Rogan WJ. Misclassification and the design of environmental studies. *Am J Epidemiol*. 1979; 109: 607– 16.

39. Jurek AM, Greenland S, Maldonado G. 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*. 2008; 37: 382– 5.

40. Fosgate GT, Hird DW, Read DH, Walker RL. Risk factors for Clostridium piliforme infection in foals. *J Am Vet Med Assoc*. 2002; 220: 785– 90.

41. McConnell EJ, Rioja E, Bester L, Sanz MG, Fosgate GT, Saulez MN. Use of near-infrared spectroscopy to identify trends in regional cerebral oxygen saturation in horses. *Equine Vet J*. 2013; 45: 470– 5.

42. Hong SW, Shim JK, Choi YS, Kim DH, Chang BC, Kwak YL. Prediction of cognitive dysfunction and patients' outcome following valvular heart surgery and the role of cerebral oximetry. *Eur J Cardiothorac Surg*. 2008; 33: 560– 5.

43. Kass PH, Greenland S. Conflicting definitions of confounding and their ramifications for veterinary epidemiologic research: collapsibility vs comparability. *J Am Vet Med Assoc*. 1991; 199: 1569– 73.

44. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003; 14: 300– 6.

45. Howards PP. An overview of confounding. Part 2: how to identify it and special situations. *Acta Obstet Gynecol Scand*. 2018; 97: 400– 6.

46. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. *BMC Med Res Methodol*. 2008; 8:70.

47. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999; 10: 37– 48.

48. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995; 82: 669– 88.

49. Simpson EH. The Interpretation of Interaction in Contingency Tables. *J R Stat Soc Series B Stat Methodol*. 1951; 13: 238– 41.

50. Kyriacou DN, Lewis RJ. Confounding by indication in clinical research. *JAMA*. 2016; 316: 1818– 9.

51. Navarro MA, Uzal FA. Pathobiology and diagnosis of clostridial hepatitis in animals. *J Vet Diagn Invest*. 2020; 32: 192– 202.

52. Swerczek TW. Tyzzer's disease in foals: retrospective studies from 1969 to 2010. *Can Vet J*. 2013; 54: 876– 80.

53. Howards PP. An overview of confounding. Part 1: the concept and how to address it. *Acta Obstet Gynecol Scand*. 2018; 97: 394– 9.

54. Joone CJ, Bertschinger HJ, Gupta SK, Fosgate GT, Arukha AP, Minhas V, et al. Ovarian function and pregnancy outcome in pony mares following immunocontraception with native and recombinant porcine zona pellucida vaccines. *Equine Vet J*. 2017; 49: 189– 95.