INTRODUCTION

In animal sciences, similarly to many other research areas and disciplines, the central goal of scientific investigations often refers to inferring causal effects. For example, livestock nutritionists, reproductive physiologists, and immunologists study environmental factors that affect growth, development, reproduction, and susceptibility to disease, such as the effect of different dietary nutrients and their concentrations on growth in meat animals, light source and lighting program on sexual maturity and reproductive performance of birds, water temperature and oxygen concentrations on fish production, and hygiene management on disease incidence in dairy species.

A common approach to infer causal effects uses randomized experiments in which subjects (e.g., animals, in the case of animal and veterinary sciences) are randomly distributed into experimental groups, each of which is then assigned to a specific treatment level, also at random. Such controlled experiments are a powerful tool for causal inference, which allow not only for testing treatment effects but also for estimating their magnitude (Fisher, 1971).

Controlled experiments, however, are not always feasible. For example, in many circumstances randomization is not possible, due to legal, ethical, or logistical constraints (Rosenbaum, 2010). Another drawback is when a randomized trial is possible only within tightly controlled experimental conditions, for example to study the effect of a certain disease on production, reproduction, or other health traits. Such experiments, however, have an important drawback for translational research. In general, as experimental settings com-
commonly used in experiment stations do not reflect the real conditions found in livestock production, their results may not be directly applicable to commercial herds.

In this context, abundant data are routinely collected in commercial herds, either for breeding purposes, health control, or general herd management decisions, which can be potentially explored to try to study causal effects of environmental and management factors (e.g., housing, diets, management practices, hormonal protocols) in livestock production, well-being, and product quality as well as how various phenotypic traits are interconnected with each other. In fact, many authors have used such available data in analyses that did not involve causal information, for example, to estimate disease incidence rates (Bartlett et al., 1986b,c), to infer genetic and phenotypic associations between traits (Bigras-Poulin et al., 1990; Zwald et al., 2004), or to develop predictive models for economically important traits (Caraviello et al., 2006; Giordano et al., 2011). Conversely, some authors have also used field data to study causal relationships between variables (Erb et al., 1981; Curtis et al., 1985). However, it should be stressed that inferring causal effects from observational data is complex due to potential confounding effects (Rubin, 1974; Rosenbaum, 2002), such that careful analyses using specific statistical and data mining techniques as well as different sets of assumptions are required. Nonetheless, although virtually unknown in the agricultural research community, such methods are available (Shipley, 2002; Pearl, 2009) and have been used in many fields, for example, in human medicine where randomized trials are frequently not feasible (Rosenbaum, 2010).

In this context, this paper reviews and discusses the analysis of observational data using field-recorded information and its potential utility in the study of causal effects. It is our postulation that there is much to be learned from such data and that the methods potentially applicable for this task could be used either to explicitly investigate causal relationships between variables or simply to generate hypotheses for further investigation using controlled experiments or additional field-recorded data. In the next 2 sections a brief discussion on the distinction between association and causation and the basics of randomized experiments are presented. Next, a discussion on the importance of observational field data on animal agriculture research is illustrated with some examples of application. Subsequently, the task of inferring causal effects from observational data is visited, and a final section is presented with concluding remarks.

**Association vs. Causation**

A well-known proverb concerning scientific inquiries and statistics states that “correlation does not imply causation.” This is a very important and factual statement that emphasizes that knowledge of correlation between 2 variables is not sufficient to describe how they are causally related. Different types of causal associations could be sources of correlation between 2 variables or to develop predictive models. Conversely, some authors have also used such data and that the methods potentially imply causation.” This is a very important and factual statement that emphasizes that knowledge of correlation between 2 variables is not sufficient to describe how they are causally related. Different types of causal associations could be sources of correlation between 2 variables or to develop predictive models. Conversely, some authors have also used such data and that the methods potentially imply causation.” This is a very important and factual statement that emphasizes that knowledge of correlation between 2 variables is not sufficient to describe how they are causally related. Different types of causal associations could be sources of correlation between 2 variables or to develop predictive models. Conversely, some authors have also used such data and that the methods potentially imply causation.” This is a very important and factual statement that emphasizes that knowledge of correlation between 2 variables is not sufficient to describe how they are causally related. Different types of causal associations could be sources of correlation between 2 variables or to develop predictive models. Conversely, some authors have also used such data and that the methods potentially imply causation.” This is a very important and factual statement that emphasizes that knowledge of correlation between 2 variables is not sufficient to describe how they are causally related. Different types of causal associations could be sources of correlation between 2 variables or to develop predictive models. Conversely, some authors have also used such data and that the methods potentially imply causation.” This is a very important and factual statement that emphasizes that knowledge of correlation between 2 variables is not sufficient to describe how they are causally related. Different types of causal associations could be sources of correlation between 2 variables or to develop predictive models. Conversely, some authors have also used such data and that the methods potentially imply causation.” This is a very important and factual statement that emphasizes that knowledge of correlation between 2 variables is not sufficient to describe how they are causally related. Different types of causal associations could be sources of correlation between 2 variables or to develop predictive models. Conversely, some authors have also used such data and that the methods potentially imply causation.” This is a very important and factual statement that emphasizes that knowledge of correlation between 2 variables is not sufficient to describe how they are causally related. Different types of causal associations could be sources of correlation between 2 variables or to develop predictive models. Conversely, some authors have also used such data and that the methods potentially
However, the notion that “correlation does not imply causation” has sometimes been taken to the limit, in which it is believed that there is no way of learning anything about causal relationships on the basis of observational data from a set of correlated variables. This is the dominant thinking in the field of agricultural sciences, in which causal inference has been largely inferred from randomized experiments, and observational data is deemed as having limited value in this regard. Nonetheless, there is extensive literature on another school of thinking, which claims that there is much more to be learned from observational data than simply correlations and covariances among variables (Spirtes et al., 2000; Shipley, 2002; Pearl, 2009). As underscored by Tufte (2003), although it is true that correlation is not causation, simply stating their nonequivalence neglects information about their relationship. Tufte (2003) suggests then that instead of “correlation does not imply causation,” we use “empirically observed covariation is a necessary but not sufficient condition for causality” or “correlation is not causation but it sure is a hint.” In this paper, we intend to contribute to the introduction and spread of such methods in the field of animal and veterinary sciences, such that we can better explore available data routinely collected in commercial herds.

**Randomized Experiments**

Before we discuss statistical and computational methods specifically useful for the analysis of observational data, here we briefly review the principles of randomized experiments as well as some of their drawbacks. A widely used strategy to infer causal effects refers to the use of controlled experiments in which subjects (animals, for example) are randomly distributed into experimental groups, each of which is then assigned to a specific treatment, also at random. Such experiments are called randomized trials; in the biomedical literature they are referred to as randomized clinical trials. Additional discussion on randomized experiments can be found, for example, in Cochran and Cox (1992) and Montgomery (2012). For example, to study the effect of bST administration on MY in dairy cattle, a homogeneous group of cows, in terms of breed, age, lactation, and any other factor deemed to affect MY or to interact with the bST effect, can be split into 2 sets, and a control (i.e., no bST administration) and a treatment (i.e., bST administration) regimen can be randomly assigned, one for each group. The MY of cows in both groups is then measured and the average performance of the 2 experimental groups can be compared using appropriate statistical techniques. For example, under the assumption of normally distributed data, a traditional $t$ test can be used to test the null hypothesis (i.e., no difference between mean MY in the 2 groups) against an alternative (single- or 2-tailed) hypothesis of difference between the 2 groups. If a statistically significant difference in mean MY is detected between the 2 groups, such a difference can be attributed to the effect of bST administered to the cows in the treatment group.

The association between bST and MY in this case can be interpreted as a causal effect of bST supplementation on MY based on the premise that the only difference between the 2 experimental groups is the bST itself and that through randomization of cows to each experimental group, all factors contributing to animal-to-animal variability are averaged out. Such controlled, randomized experiments allow not only for testing for a causal effect of a specific treatment (i.e., bST in this case) but also for estimating such effects.

The reasoning behind randomized experiments can also be articulated in terms of causal diagrams. If the goal is to study the causal effect from $x$ to $y$ and the causal structure involving these variables is as presented in Fig. 1A, the observed association between them could not be used to make inferences because there is an extra source of association due to an unmeasured variable, $z$, that affects both variables. However, in a randomized experiment, the levels of $x$ are randomly assigned by the researcher so that variable $z$ no longer has a contributing effect. Under this scenario, the underlying causal diagram would be as presented in Fig. 1B, in which the edge between $x$ and $z$ is removed. Because individuals are assigned at random to groups under specific levels of $x$, $z$ is averaged out and the only remaining source of association between $x$ and $y$ is the causal effect from the former to the latter. Therefore, causal inferences are drawn from associations detected under this trial. Notice also that if there were no causal relationships between $x$ and $y$ (i.e., if there was no connection between them in the causal structure in Fig. 1A), there would be no sources of associations under the randomization of $x$ so that any association would be expected to vanish. That result would be interpreted as lack of evidence against the hypothesis that $x$ does not affect $y$. Finally, if the underlying causal structure in Fig. 1A was such that the

![Figure 1. Causal structures involving variables $x$, $y$, and $z$ under no external interventions (panel A) and under the external intervention from which the value of $x$ is randomly assigned (panel B).](image-url)
edge between $x$ and $y$ was directed toward $x$ instead of toward $y$, the randomized assignment of the values of $x$ would remove this edge as well. Therefore, the association between $x$ and $y$ would also be expected to vanish in this case, and $x$ would be declared as not having an effect on $y$. More details about interventions and models based on causal graphs are presented subsequently.

Experimental designs and the many statistical techniques used in their analysis, such as ANOVA, go back to the work of Sir Ronald Aymer Fisher (see Fisher, 1971) in agricultural experimentation, and they are widely used in animal sciences and many other fields. Nowadays complex, multifactorial experiments are used to concurrently study multiple factors with multiple levels each and their interaction or joint effects on outcomes of interest. Likewise, sophisticated parametric and nonparametric statistical techniques are available for the analysis of randomized trials, including multivariate approaches to study multiple response variables simultaneously, longitudinal data, nonlinear responses, and binary and counting variables, among others.

However, randomized experiments are not always workable. In many situations randomization is not possible due to legal, ethical, or logistical constraints. For example, to study the effect of smoking or alcohol consumption on human health via a randomized clinical trial would require randomizing individuals to smoking or alcohol consumption groups, which would be evidently unfeasible. Although related experiments could be performed with other animal species, for example with model organisms, translating the results to humans would still require extensive investigation.

Another example would relate to the study of causal relationships among phenotypic traits. For example, to study the causal effect of obesity (e.g., body mass index [BMI]) on hormone profile changes using controlled experiments, individuals would need to be randomized to obese or nonobese groups and then have their BMI changed according to group assignment. However, increasing BMI of individuals allocated to the obese group or, Likewise, decreasing BMI of individuals allocated to the nonobese group would require a number of external interventions, such as changes to physical exercise programs or calorie intake, which could also directly affect the hormonal profile of the individuals, confounding the study in terms of the direct effect of fat deposition on hormone concentrations. Similar difficulties would be imposed in investigations involving other sets of phenotypes, such as studying the effect of weaning weight on first calving interval in cattle or the effect of age at sexual maturity on egg production in chickens.

The impossibility of carrying out a randomized experiment in many scenarios certainly complicates the test and estimation of causal effects. Nonetheless, even in such situations it is still possible to explore potential causal effects using observational studies, although very specific statistical and data mining methodologies are necessary, and considerable caution should be taken when analyzing and interpreting observational data. A more detailed discussion on these topics is presented subsequently.

In addition to the potential unfeasibility of performing randomized experiments in some circumstances, experimental settings used for livestock research in general do not reflect the real conditions found in commercial herds. Therefore, although experiments performed in research station herds are still an extremely important component of a series of studies necessary for a better understanding of causal relationships between variables, the results obtained from experiments with highly controlled conditions are often not directly translated to real world conditions. It is quite common that effects inferred from data recorded in experimental settings represent an overestimation of the same effects as assessed in commercial farms and, more importantly, that the effect of a specific factor is highly dependent on the levels of other factors (i.e., there exists interaction between factors), which may be variable in commercial settings but were held constant in the controlled experiment. Therefore, translational research inevitably involves additional experiments within commercial farm conditions to validate results found in controlled trials and to reestimate effects and parameter values reflecting the ultimate use of the basic research findings. Randomized experiments can then be performed also in commercial herd settings and involving extra factors (such as different breeds and different management practices) to be studied jointly using multifactor experiments and on realistic farm conditions (Tempelman, 2009; Engstrom et al., 2010). This would be a top-down approach in which basic research questions are first investigated in controlled conditions, as an initial stage, and then results are validated in commercial settings.

Another interesting approach in livestock research is to take advantage of the wealth of data routinely collected in commercial herds, either for breeding purposes, health control, or general herd management decisions, and explore it to try to better understand the causal effects of environmental and management factors on livestock production, well-being, and product quality. Such data are available for a number of phenotypic traits, including production, reproduction, and product quality traits, and for a myriad of environmental factors, including both producer-managed factors (such as diet, housing, and management practices) and environmental factors (such as temperature, humidity, soil fertility, water abundance, etc.). In the next section we present some examples of studies using observational field data for different purposes, including also research on causal effects between variables.
Exploring Observational Field Data

Data routinely collected on commercial farms for different species and breeds are essential for monitoring health and production of herds as well as for decision making within management practices. These data have also been used extensively for quantitative genetic analysis (e.g., estimation of heritabilities and genetic correlations) and genetic improvement of animal populations. In addition, such available field data have been used for other research purposes and goals, as for example to estimate disease incidence rates and their interrelationships, to develop predictive models for economically important traits, or even to study causal relationships between variables.

An example of field data use was presented by Kaneene and Hurd (1990), who discussed the use of producer-recorded health data from the National Animal Health Monitoring System in Michigan, which was designed to collect and curate data for use in estimating disease frequencies in dairy cattle. Randomly selected dairy herds were visited periodically to collect management, disease, inventory, production, preventive treatment, financial, and any other relevant data using custom-developed worksheets, with strict data quality control devices used. In another example of field data use in dairy cattle, Lyons et al. (1991) performed genetic analysis of 22 health related traits using data partially gathered from dairy producers in Wisconsin, Minnesota, and Iowa. Using traditional multtrait genetic models, they inferred genetic and phenotypic correlations between production and health traits. These 2 examples, however, were still relatively small studies, based on field data collected from farm paper records.

An alternative to allow for broader studies is to make use of data from on-farm computer management systems, which has been recently advocated as an effective and low-cost source of health trait data in dairy cattle (Bartlett et al., 1986a; Parker-Gaddis et al., 2012). In the United States, studies using such data have been published, for example, by Bartlett et al. (1986b,c) and Zwald et al. (2004).

Bartlett et al. (1986b,c) examined the incidence and epidemiological characteristics of metritis and of cystic follicular disease using data collected from 22 Michigan dairy herds participating in a computerized herd health program. The authors also estimated the extra cost associated with a cow treated for metritis as well as the economic loss for the average lactation with cystic follicles.

Zwald et al. (2004) investigated the feasibility of genetic selection for health traits in dairy cattle using data recorded in on-farm herd management software programs. Health traits included displaced abomasum, ketosis, mastitis, lameness, cystic ovaries, and metritis, which were collected during a period of 3 yr from herds using Dairy Comp 305 (Valley Agricultural Software, Tulare, CA), DHI-Plus (DHI Computing Service, Inc., Provo, Utah), or PCDART (Dairy Records Management Systems, Raleigh, NC) herd management software programs. After quality control, data from about 210 to 430 herds and including around 50,000 to 105,000 cows total, depending on the health trait, were used for assessing incidence rates and estimating heritabilities. As a general conclusion, the authors indicated that data from on-farm recording systems could be successfully used for genetic selection against common health disorders in dairy cattle.

Some other authors investigated associations between disease traits, such as Bigras-Poulin et al. (1990), who performed an observational study with 34 Holstein herds from 7 southwestern Ontario counties to evaluate common clinical diseases in terms of frequency of occurrences, time to first diagnosis, and associations among such diseases. The data consisted of 2,204 lactations, with data on incidence of milk fever, udder edema, fetal membranes retained, ketosis, vaginitis, metritis, mastitis, teat injuries, foot and leg problems, ovarian cysts, and abortions. Results indicated that cows younger than 3 yr of age were at very low risk of experiencing milk fever. In addition, the author indicated that a potential causal association was found between retained placenta and each of metritis and ovarian cysts as well as between teat injuries and mastitis and between milk fever and ketosis.

Other studies used field data to develop predictive models. For example, Caraviello et al. (2006) used machine learning algorithms to identify factors associated with reproductive performance of lactating Holstein cows on large dairy farms and developed a classification algorithm for first-service conception rate and pregnancy status at 150 d in milk. The authors used production and reproductive records from 153 farms obtained from on-farm DHI-Plus, Dairy Comp 305, or PCDART herd management software. Survey data regarding management, facilities, labor, nutrition, reproduction, genetic selection, climate, and milk production were also available as well as BCS and environmental temperature data. The edited data consisted of about 31,000 lactation records from 14,800 cows and 317 regressors (or predictor variables) for first-service conception rate and about 17,500 lactation records from 9,500 cows and 341 regressors for pregnancy status at 150 d in milk.

With a slightly different goal, Giordano et al. (2011) used field data from dairy herds to develop an analysis tool to obtain the net present value (US$/cow per year) of different reproductive management programs. The model used productive, reproductive, and economic data as input to simulate farm conditions to account for all factors deemed to be related to reproductive management that increase costs and generate revenue. A case study was presented in which the model was used to compare 3 different reproductive management strate-
gies using data from a commercial farm, and sensitivity analysis was used to assess the effect of varying specific reproductive parameters.

More specifically on the investigation of potential causal relationships between variables, Erb et al. (1981) considered 5 disease traits (defined as “endogenous variables”), namely dystocia, retained placenta, metritis, cystic follicle, and luteal cyst. The authors used a path analysis model, which also included 7 other variables, defined as “exogenous variables”: age, previous and current breed class average for milk, previous and current days in milk, days dry, and calving interval. The authors used time sequence information as well as additional prior knowledge to postulate a general causal structure connecting the variables, from which 6 competing models were derived and compared using a chi-square-based test of adequacy. Partial regression coefficients (or path coefficients), which parameterize arrows in a path analysis graph, were then estimated and tested against the null hypothesis of inexistence of each specific arrow connection. The authors reported a total of 40 bivariate relationships, with 34 classified as allegedly causal. Highlighted results were that breed class average, days in milk, and dry period in the previous lactation had little effect on disease. Some relationships were detected among the diseases, which were assumed to be causal, for example, the effect of retained placenta on metritis. Also, with the exception of dystocia, all of the diseases were allegedly causative of increased calving interval, breed class average for milk, and days in milk.

A similar study was presented by Curtis et al. (1985), who used path analysis and logistic regression to model direct and indirect relationships among clinical periparturient retained placenta, metritis, veterinary-assisted dystocia, uncomplicated and complicated ketosis, left displaced abomasum, parturient paresis, mastitis, and estimated nutrient intakes (protein, calcium, phosphorus, and energy; coded into terciles) in the last 3 wk of the dry period. Data were from 1,374 multiparous Holstein lactations in 31 commercial herds in central New York. An initial underlying path (or graph) was constructed considering paths that were thought to be plausible or that had been previously documented. Partial regression coefficients and their derived conditional odds ratios were then sequentially tested against a null hypothesis using a stepwise multiple logistic regression procedure. Estimated odds ratios indicated that parturient paresis increased the incidence of veterinary-assisted dystocia, retained placenta, complicated ketosis, and clinical mastitis. In addition, retained placenta, left displaced abomasum, and parturient paresis were found to directly increase the risk of complicated ketosis. Greater terciles of estimated energy intake in the last 3 wk of the dry period decreased the risk of veterinary-assisted dystocia and left displaced abomasum whereas greater terciles of estimated protein intake decreased the risk of retained placenta and uncomplicated ketosis. Estimated nutrient intakes were directly related to subsequent metabolic disorders and directly and indirectly related (mediated by metabolic disorders) to reproductive disorders. In conclusion, the authors indicate that greater intakes of protein and energy in the last 3 wk of the dry period might reduce the incidence of metabolic and reproductive disorders. However, they were careful to point out that specific recommendations as to the amounts and types of feed could not be made from their analyses.

Another example with path analysis was presented by van Dorp et al. (1999), who used data recorded in a herd health management system from 32 registered Holstein dairy herds from British Columbia to assess the frequencies of various diseases and to study the effect of herd, age, year, and season as well as the interrelationships between diseases. An informal path model of disease interrelationships was adopted, from which 2 independent pathways were inferred: one started by udder edema and the other by milk fever. On those paths, udder edema was directly associated with mastitis occurrence, metritis, and cystic ovaries. Mastitis in early lactation increased the risk of mastitis and cystic ovaries in later lactation. Milk fever was directly related to displaced abomasum, which increased the risk of footrot. In general, the authors concluded that diseases that occurred in early lactation tended to increase the risk of other diseases later in lactation.

More recently, Parker-Gaddis et al. (2012) assessed the plausibility of producer-recorded health data in the United States. They also examined putative relationships among common health events, building path diagrams developed using odds ratios calculated from logistic regression techniques. Logistic models were built based on associations observed between common health traits as well as the timeline separating disease events; that is, health events occurring on average earlier in life were used as predictors for the health trait of interest. Their general conclusions were that data collected from on-farm computer management systems might provide an effective and low-cost source of health information. However, they properly indicated that a more detailed study would be necessary to accurately infer causal relationships between those health traits.

As illustrated with the examples discussed previously, many authors have been using field-recorded data for different research goals, some of which including also the investigation of causal effects. In such cases, path analysis techniques are generally used, with an initial causal structure derived from prior knowledge, either from previous studies or from temporal sequence. A related technique, which allows for the modeling of genet-
ic relationships between animals together with putative causal links between phenotypic traits, refers to structural equation models (StEM; Wright, 1921; Haavelmo, 1943) embedded within a mixed model methodology (Gianola and Sorensen, 2004; Wu et al., 2010). A review and discussion of StEM applied to livestock data is presented by Rosa et al. (2011). This approach has been used by many authors, working with different species and traits, some of which are discussed below.

In 2006, de los Campos et al. (2006) presented one of the first applications of StEM to study recursive or simultaneous effects between traits within a quantitative genetics mixed effects model. They studied the relationship between MY and somatic cell score (SCS) in dairy goats using data related to repeated measurements in each half of the udder of the animals. Results indicated a negative effect of SCS on MY, with not much evidence in favor of a dilution effect. In addition, the authors found simultaneity of effects between SCS from the left and right halves of the udder.

Varona et al. (2007) performed an analysis of litter size and average piglet BW at birth in Landrace and Yorkshire pigs using a standard 2-trait mixed model and a recursive mixed model. The authors, however, were more interested in comparing models in terms of their posterior predictive performance and not exactly of the interpretation of putative causal links between traits. Finally, as an example with dairy cattle, de Maturana et al. (2009, 2010) used Gaussian-threshold StEM with heterogeneous structural coefficients to explore biological relationships between gestation length (GL), calving difficulty (CD), and perinatal mortality (or stillbirth [SB]) in Holsteins. An acyclic model was assumed, which included a recursive effect from GL to the liabilities (latent variables) of CD and SB and from the liability of CD to that of SB. The results indicated that gestations approximately 274 d in duration (i.e., ~3 d less than the average) led to the least CD and SB levels and confirmed the existence of an intermediate optimum of GL with respect to these traits.

The studies discussed previously, using either path analysis or StEM, attempted to go one step further on the investigation of causal effects, as compared with previous analyses restricted to estimation of correlation and covariances or aimed toward predictions. However, they all rely on prior knowledge of the causal structure, so they are in some sense more appropriate for inferring the magnitude of postulated causal effects or at maximum to test some local, specific causal links within an overall hypothesized causal structure. An extension of such methods, which performs a full search of plausible causal structures and accounts for confounding effects resulting from additive genetic correlations between traits, has been proposed by Valente et al. (2010). Valente et al. (2011) presented the first application of this approach with real data, working with a data set on 5 production and reproduction traits in European quail: birth weight, BW at 35 d of age, age at first egg, number of eggs laid from 77 to 110 d of age, and average egg weight. The authors focused their discussion on the challenges and difficulties resulting from applying this method to field data. They pointed out that coupling prior knowledge with the output of the analysis using the inductive causation (IC) algorithm (Verma and Pearl, 1990; Pearl, 2009) allowed for further understanding regarding phenotypic causal structures compared with standard mixed effects StEM applications.

In summary, the usefulness and importance of field data for animal agriculture research is unquestionable. Such data, routinely collected from commercial herds for many livestock species, can be used even for the investigation of potential causal relationships between disease, reproduction, and production traits as well as for studying how they are affected by environmental and management factors. However, inferring causal effects from observational data is complex due to potential confounding effects, such that careful analyses using specific statistical and data mining techniques are required.

Nonetheless, although virtually unknown in the agricultural research community, such methods are available and have been extensively used in many fields (Spirtes et al., 2000; Shipley, 2002; Pearl, 2009), for example, in human medicine in which quite often randomized trials are not feasible (Rosenbaum, 2010). In the next section we provide a brief discussion on the task of inferring causality from observational data.

**Inferring Causal Effects From Observational Data**

The task of inferring causal effects from observational data is challenging and polemic. One reason for the controversy is that inferring such information from observational studies always requires additional assumptions relative to standard statistical inferences that do not embody causal meaning (e.g., estimating correlations, regression coefficients). However, the information provided by these 2 types of inference is very different. Statistical information essentially describes how plausible a given event (which may be multivariate) is. On the other hand, causal information describes how the value of one variable is affected by the value of other variables. The practical implication is that causal information regarding a set of variables allows one to predict the outcomes of external physical interventions on the causal network, which may be very useful for livestock management. Such predictions could not be obtained from a joint distribution of such a set of variables (Spirtes et al., 2000; Pearl, 2009). As a consequence, it is not fair to
compare the strength of required sets of assumptions for both types of inferences as if they were alternatives for obtaining the same information.

Different models, such as Bayesian networks (BN) or StEM, are used to express causal relationships among multiple traits. Fitting these models allows inferring causal effects conditionally on a causal structure given a priori. This type of analysis permits predicting the outcomes of external interventions from observational data and some causal assumptions represented by the causal structure of the model. In the case of StEM, the causal network is represented by a directed graph, and each family of nodes involving parents and a child in the graph represents a structural equation in which the variable on the left-hand side (LHS) is determined by a function, generally constructed as linear, of the variables on the right-hand side (RHS). Therefore, the equal sign in these equations represents the asymmetrical relationship defined as “is determined by,” which differs from its usual meaning in standard equations. The interpretation is that if the variables in the RHS assumed some arbitrary specific values, the quantity in the LHS would be defined by a function of the variables in the RHS (Pearl, 1995, 2009). The causal meaning of structural equations is essentially the same as that presented in the formulation of Rubin (Rubin, 1974; Rosenbaum and Rubin, 1983) although these authors articulate these relationships in a different manner using concepts such as potential outcomes and mechanisms of treatment assignment.

The causal interpretation derived from fitting a model, such as StEM or BN, depends not only on statistical assumptions but also on causal assumptions. For example, inferences regarding causal associations obtained by fitting a BN require assuming that the causal structure among variables is acyclic and that all the variables that have causal effects on 2 or more variables contained in the BN are already in the BN. Identical assumptions could be applied when using StEM, and they are translated into using acyclic causal structures and constructing the residual covariance matrix as diagonal, which is sufficient to guarantee any recursive StEM to be identifiable from data. However, this assumption is not necessary for StEM, which could present cycles and account for associations due to hidden variables, represented by residual covariances. Nonetheless, identifiability is not guaranteed if these features are allowed and, therefore, must be verified (Spirtes et al., 2000; Gianola and Sorensen, 2004).

In general, fitting one of these models or using other strategies for making inferences about the causal relationships between traits (which will be addressed subsequently) requires an important assumption. Specifically, it is assumed that the associations between measured variables, or at least between some of them, are not confounded by unknown variables that present causal effects on 2 traits or more. Alternatively, it is at least assumed that the resulting hidden paths that confound the causal inference are partially known and that measurements of variables contained in them are possible. Here the criticisms of causal inference based on observational data argue that we can seldom be completely confident that there are no sources of confounding bias or that by controlling for such paths that might be a source of confounding, there are no further sources of confounding (Cox and Wermuth, 1995). This is a reasonable argument that has an impact on observational studies, but it does not imply the need for abandoning the task of learning causality altogether. This criticism indicates that the conclusions about a causal effect cannot bear the same level of certainty involved in declaring 2 variables as probabilistically associated.

However, there is a wide range of possible degrees of knowledge or confidence between absolute certainty and absolute ignorance. The formal impossibility to reach absolute certainty regarding inference of causal relationships does not mean we are compelled to remain in absolute ignorance. Additionally, the purely statistical approach is not exactly an alternative, because as previously discussed, causal and statistical information present distinct meanings and usefulness. It seems reasonable to interpret results of causal inferences with some degree of caution. Nonetheless, such studies may enlighten some important aspects of relationships among variables, for example, pointing out some interesting variables to be measured in a future study of a possible causal association or indicating which causal associations are more likely to be true and deserve further research and which randomized trials would be more informative for understanding a complex causal network under study.

As discussed previously, the ability to predict the effect of interventions is one of the most important features provided by causal information. Different types of interventions can be predicted from causal models. The most common type of intervention is setting externally the value of a variable in the network. The effect of this kind of intervention may be represented by wiping out equations, also referred to as model surgery or mutilation (Pearl, 1995, 2009). For example, consider the following StEM:

\[
\begin{align*}
  y_{i1} &= \mu_1 + e_{i1} \\
  y_{i2} &= \mu_2 + \lambda_{21} y_{i1} + e_{i2} \\
  y_{i3} &= \mu_3 + \lambda_{32} y_{i2} + e_{i3} \\
  y_{i4} &= \mu_4 + \lambda_{42} y_{i2} + e_{i4} \\
  y_{i5} &= \mu_5 + \lambda_{53} y_{i3} + \lambda_{54} y_{i4} + e_{i5},
\end{align*}
\]

which represents a structure such as that depicted in Fig. 2A. Wiping out equations allows one to infer the outcome in...
of this action without data recorded under manipulation. For example, the effect of setting \( y_3 \) to an arbitrary constant \( c \) would be represented by

\[
\begin{align*}
  y_{i1} &= \mu + e_i \\
  y_{i2} &= \mu_2 + \lambda_{21} y_{i1} + e_i \\
  y_{i4} &= \mu_4 + \lambda_{42} y_{i2} + e_i \\
  y_{i5} &= \mu_5 + \lambda_{53} + \lambda_{54} y_{i4} + e_i
\end{align*}
\]

which can be graphically represented as in Fig. 2B. Here we notice that although \( y_3 \) is closely connected to \( y_2 \) and \( y_5 \), setting \( y_3 \) to different values would only have consequences in the value of \( y_5 \), but the value of \( y_2 \) would remain unaltered, as described by the structural equations. That is, \( y_5 \) is defined as a function of \( c \), but \( y_2 \) is not. The values of \( y_1 \) and \( y_4 \) are also unaffected by interventions in \( y_3 \). The effect of other types of external interventions can be computed similarly, such as the effect of intervening in the distribution of a variable without modifying the causal relationships in the network, also known as soft or parametrical intervention (Eberhardt and Scheines, 2007; Eberhardt et al., 2010), and the effect of changing the functional relationship between nodes (Pearl, 2009).

Inferring the effects of these interventions relies on the knowledge of causal relationships among traits. In cases in which causal links between variables are not assumed to be known, one may use algorithms that allow exploring spaces of causal structures, such as the IC algorithm (Verma and Pearl, 1990), the Spirtes-Glymour-Scheines (SGS) algorithm (Spirtes et al., 1990), and the Peter and Clark (PC) algorithm (Spirtes and Glymour, 1991) among others. These algorithms are based on a series of causal assumptions (Spirtes et al., 2000), from which the causal sufficiency assumption is likely the strongest. Nevertheless, some learning is possible even if we do not assume causal sufficiency. For example, the Fast Causal Inference (FCI) algorithm (Spirtes et al., 2000) drops the causal sufficiency assumption although it results in a more complex output graph. Prior information about the involved variables (e.g., temporal information or causal associations already learned from randomized experiments) may aid in learning causal associations with fewer assumptions and consequently improving overall inferences (e.g., Valente et al., 2011).

Representation of causal networks with directed graphs allows efficient expression of the assumptions underlying such networks. The expected probabilistic conditional independencies that follow from causal assumptions are given by d-separations (Pearl, 1988, 2009) in the graph. The directed graphs permit checking whether the assumptions are sufficient for estimating a target quantity (e.g., a single causal effect). For example, if the goal is to study the causal relationship between 2 traits in a setting where features of the causal structure that encompass these 2 traits are known but measuring all traits is unfeasible, one issue is verifying what the conditions are that make the target causal effect identifiable from observational data (Pearl, 1995).

As an example of this application, consider the possibility that the inference of the causal effect between 2 variables, \( x \) and \( y \), is hindered by confounding bias generated from paths involving unmeasured variables that are additional sources of correlations between them, as paths \( x \leftarrow z_2 \rightarrow y \) and \( x \leftarrow z_2 \rightarrow z_3 \rightarrow y \) in Fig. 3. Rosenbaum and Rubin (1983), using the concept of potential outcomes to study causal models, formulated rules to declare a set of concomitant variables \( Z \) as sufficient to be considered in the analysis to allow for the identification of the target causal effect from observational data. Pearl (1993) formulated these rules in terms of graphical tests, coming up with the so-called back-door criterion: The total causal effect of \( x \) on \( y \) in a causal diagram \( G \) can be computed from a set of measured traits \( S \) (additional to observing \( x \) and \( y \)) such that 1) no member of \( S \) is a descendant of \( x \) and 2) \( S \) d-separates \( x \) from \( y \) in the subgraph formed by deleting from \( G \) all arrows emanating from \( x \).
In the causal structure given in Fig. 3, to infer the total causal effect of \( x \) on \( y \), it would be sufficient to study the associations between the 2 nodes conditionally on \( S = \{z_3\} \), which blocks all the back-door paths. One could then make inferences about this causal relationship (represented by the coefficient \( \lambda \)) by fitting the following regression model:

\[
y = \lambda x + \beta z_3 + e.
\]

Another strategy for inferring causal effects between 2 variables is by studying a broader causal association network containing them. A well-known example of this strategy is the use of instrumental variables. For the classical setting of Fig. 4, the coefficient that represents the target causal effect \( \lambda \) is determined from the association between \( z \) and \( x \) (\( \beta \)) and \( z \) and \( y \) (\( \beta \lambda \)), that are both identifiable, so that \( \lambda = \beta \lambda / \beta \). This strategy of inferring causal effects does not work, however, for causal models with nonlinear (or unspecified) causal functions.

Conversely, causal relationships studied on the basis of the back-door criterion do not require assuming any parametric description of causal relationships. As always, causal assumptions are required to estimate the effect on \( y \) originated from manipulating \( x \), but one can verify for any setting if and how such a task is possible from observational data. The method of verification combines the information in a causal diagram (with back-door paths represented with bidirected arcs or latent variables) and the use of a set of inference rules given by Pearl (1995). This is performed without any assumption regarding the form of the causal function among traits. By performing deduction based on these rules, one may transform the sought causal effect from intervention into an equivalent expression exclusively involving standard probability functions based on observational data. For example, consider that one is interested in studying the causal effect of variable \( x \) on variable \( y \) (here represented as \( p(y \mid \bar{x}) \), in which \( \bar{x} \) means that \( x \) is set from external manipulation) on the structure depicted in Fig. 5. In this structure, there is a path involving unobserved variables responsible for some degree of confounding bias represented by a bidirected arc, and the target causal effect is mediated by \( z \). It can be proven that the target information can be reduced into

\[
p(y \mid \bar{x}) = \sum_{x} p(z \mid x) \sum_{x'} p(y \mid x', z) p(x')
\]

which presents no variables under intervention and meets the front-door criterion for identifying causal effects (Pearl, 1995).

Notice that the application of the above-mentioned rules can make explicit some aspects that contradict the usual practices in similar studies. For the example given, notice that to identify the causal effects of \( x \) on \( y \), it is necessary to account for a variable \( z \) that mediates this effect, which is generally deemed as something forbidden in this case. Although contradicting a common practice, the causal relationship obtained in this setting relies on fewer assumptions than, for example, causal effects studied via instrumental variables, as in Fig. 4. Using instrumental variables renders the causal effect as identifiable not solely from the causal assumptions expressed by the graph but also from the linearity assumed for causal effects (Pearl, 2009). However, by using the calculus for intervention, we can check that the same setting does not allow identification if we refrain from assuming causal relationships as represented by parametric functions. On the other hand, linearity is not assumed for causal inference based on settings that meet the front-door criterion (as in Fig. 5) as well as for any setting for which identifiability of causal effects is proven to be derived from observational studies from calculus of interventions.

In summary, causal information is essential for developing methods and practices as well as for informed decision making in livestock management, given that many of them involve external interventions. Although randomized experiments are the most accepted and applied way of seeking such information, these trials are not feasible in many circumstances. This unfeasibility, however, does not make it necessary for causal information to vanish. As discussed earlier, some researchers in animal sciences have made initial steps in trying to tackle this issue, via path analysis and StEM. However, the concepts presented in this section would allow deeper understanding of causal networks and better use of this knowledge. For instance, the prediction of the consequences of external interventions through modifications
the graph would be very useful in management decisions and for development of new techniques or new drugs. Concepts involved in causal diagrams and rules of inference could be explored as a sound guide in studying how variables are causally related using observational data and describing more explicitly the meaning of the results and the assumptions undertaken. In addition, they make it possible to investigate if a specific setting allows inference of causal effects, indicating also the required conditions to make it possible if the target causal relationship is not identifiable. Surprisingly, such concepts and methodologies have been rarely used in animal agriculture research. This is even more remarkable when we consider that probably the most expensive component of livestock research is collecting data and that with on-farm recorded information the data are already available. Currently, huge amounts of data are available for many traits in many species and reflecting real commercial farm conditions. Animal science researchers should take advantage of the availability of such data and use modern causal model techniques to advance our knowledge regarding causal effects involving economically and socially important traits in livestock.

**SUMMARY AND CONCLUSIONS**

Field data, such as on-farm recorded information, is a valuable resource for research in livestock. They are virtually cost free, generated as a by-product of recording systems put in place, for example, for guiding management practices, genetic improvement of herds, or health surveillance. Such data can be used either to validate research findings from controlled experiments performed at experiment stations or as a generator of new causal hypotheses to be explored further. As emphasized in this paper, considerable caution must be taken when interest refers to the investigation of causal effects using such observational data. Nonetheless, suitable concepts and methodologies for such a task are available and have been used in other fields although they are virtually unknown in the agricultural research community. With the appropriate analysis and exploration of field data, we can advance our knowledge regarding causal relationships between variables, which in turn can enhance decision making concerning management practices in livestock.

**LITERATURE CITED**


