ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Evaluation of Causal Effects and Local Structure Learning of Causal Networks

Zhi Geng,[1] Yue Liu,[1] Chunchen Liu,[2] and Wang Miao[3]

[1]School of Mathematical Sciences, Peking University, Beijing 100871, China; email: zhigeng@pku.edu.cn, ly199125@pku.edu.cn

[2]Department of Data Mining, NEC Laboratories China, Beijing 100600, China; email: liu_chunchen@nec.cn

[3]Guanghua School of Management, Peking University, Beijing 100871, China; email: mwfy@pku.edu.cn

Annu. Rev. Stat. Appl. 2019. 6:103–24

First published as a Review in Advance on October 10, 2018

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-030718-105312

Keywords

causal inference, causal network, confounder, instrumental variable, negative control, local structure learning, surrogate paradox, Yule-Simpson paradox

Abstract

Causal effect evaluation and causal network learning are two main research areas in causal inference. For causal effect evaluation, we review the two problems of confounders and surrogates. The Yule-Simpson paradox is the idea that the association between two variables may be changed dramatically due to ignoring confounders. We review criteria for confounders and methods of adjustment for observed and unobserved confounders. The surrogate paradox occurs when a treatment has a positive causal effect on a surrogate endpoint, which, in turn, has a positive causal effect on a true endpoint, but the treatment may have a negative causal effect on the true endpoint. Some of the existing criteria for surrogates are subject to the surrogate paradox, and we review criteria for consistent surrogates to avoid the surrogate paradox. Causal networks are used to depict the causal relationships among multiple variables. Rather than discovering a global causal network, researchers are often interested in discovering the causes and effects of a given variable. We review some algorithms for local structure learning of causal networks centering around a given variable.

1. INTRODUCTION

In scientific research, the primary goal is to evaluate the causal effect of a given treatment or exposure on a given outcome or response variable. However, statistical association and correlation do not imply causation, and vice versa, especially in observational studies. On the one hand, spurious correlation between two variables may appear even if they do not have causal relationship; on the other hand, apparent independence of two variables may occur even if they have causal relationship. The well-known Yule-Simpson paradox (Yule 1903, Simpson 1951) suggests that association between the treatment and the outcome may change dramatically due to ignoring confounders that affect both of them. Hence, the paradox reveals the crucial role of identifying and adjusting for confounders in causal effect evaluation. Confounding and confounders are basic concepts in causal effect evaluation. The confounding bias is the difference between the true causal effect and its estimate. The covariates leading to the confounding bias are called confounders. Greenland et al. (1999b) and Greenland & Pearl (2011) gave overviews of these concepts. In Sections 2 and 3, we review the criteria for confounders and some methods of adjustment for observed and unobserved confounders.

In clinical trials, psychological studies, and much other scientific research, surrogates of true endpoints of interest are used owing to measurement costs, duration, or unobservability of the true endpoints. For example, CD4 cell count is used as a surrogate for survival time in randomized trials about HIV. Criteria for choosing surrogates are essential in such studies. Several criteria for surrogates have been introduced, such as the statistical surrogate (Prentice 1989), the principal surrogate (Frangakis & Rubin 2002), and the strong surrogate (Lauritzen 2004). However, the surrogate paradox presented by Chen et al. (2007) showed potential pitfalls of using such criteria: A treatment has a positive causal effect on a surrogate, and the surrogate has a positive causal effect on the true endpoint, but the treatment may have a negative causal effect on the true endpoint. In Section 4, we review the criteria for the consistent surrogate to avoid the surrogate paradox based on the knowledge of the causation or the association among treatment, surrogate and endpoint.

Causal networks are used to depict the causal relationships among multiple variables (Pearl 2009). In previous volumes of this journal, Drton & Maathuis (2017) and Heinze-Deml et al. (2018) have reviewed several approaches for structure learning of causal networks with observational data and experimental data. In Section 5, we focus on the local structure learning of causal networks around a given target variable.

In Section 6, we further discuss some issues about the adjustment for nonconfounders, testability of conditions in the criteria for the consistent surrogate, and the local structure learning approach from the perspective of sequential observational studies.

2. CONFOUNDERS AND THEIR CRITERIA

Throughout, we let X denote a treatment or exposure and Y a response or outcome variable; we use lowercase letters to denote realized values of random variables unless otherwise stated, e.g., y for a realized value of Y. We focus on a binary treatment or exposure and let $X = 1$ for treated or exposed population and $X = 0$ for control or unexposed. The Yule-Simpson paradox shows that it is necessary to identify and adjust for confounders in causal effect evaluation. Let V denote a potential confounder. Let $V \perp\!\!\!\perp Y|X$ denote that V is independent of Y conditional on X, and $V \not\perp\!\!\!\perp Y|X$ otherwise. Based on various examples in epidemiological studies, Miettinen & Cook (1981) suggested that a confounder V must satisfy the following two conditions:

1. It is predictive of risk in the unexposed population, i.e., $V \not\perp\!\!\!\perp Y|X = 0$.
2. It is distributed differently in the exposed and unexposed populations, i.e., $V \not\perp\!\!\!\perp X$.

After the work of Miettinen & Cook (1981), two classes of rigorous criteria for confounders were proposed and discussed in the literature:

1. The comparability-based criterion: A covariate is a confounder if the potential outcome distributions in the exposed population differ from those in the unexposed population by omitting the covariate
2. The collapsibility-based criterion: A covariate is a confounder if the association measure or the parameter of interest is affected by omitting the covariate (i.e., it is not collapsible over the covariate)

The comparability-based criterion exploits the potential outcome framework to define confounders. We introduce the potential outcome framework before discussing the comparability-based criterion. Following the convention, we maintain the stable unit treatment value assumption (SUTVA) that "the potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes" (Imbens & Rubin 2015, p. 10; Rubin 1980). Under SUTVA, we use Y_1 and Y_0 to denote the potential outcomes that would be observed under treated or exposed $X = 1$ and control or unexposed $X = 0$, respectively. The observed outcome Y is a realization of the potential outcome under the treatment accepted in reality, i.e., $Y = Y_x$ if $X = x$. In the potential outcome framework, causal effects are defined through comparisons of potential outcomes. Letting $Y_x(i)$ denote the potential outcome for individual i, the individual causal effect (ICE) of X on Y for i is $\text{ICE}_i = Y_1(i) - Y_0(i)$. The average causal effect (ACE) of X on Y is $\text{ACE} = \text{E}(Y_1 - Y_0)$; the ACE in the exposed population is $\text{ACT} = \text{E}(Y_1 - Y_0 | X = 1)$. For example, epidemiologists are mostly interested in the causal effect of smoking on lung cancer in smokers, not in nonsmokers. For each unit, only one of the potential outcomes can be observed.

Suppose that we are interested in the ACE in the exposed population. Then the comparability-based criterion identifies a covariate set V as confounders if the exposed population is comparable to the unexposed population conditionally on V but not after omitting V, i.e., $\text{pr}(Y_0 = y \,|\, X = 1, V) = \text{pr}(Y = y \,|\, X = 0, V)$ but $\text{pr}(Y_0 = y \,|\, X = 1) \neq \text{pr}(Y = y \,|\, X = 0)$, where $\text{pr}(\cdot|\cdot)$ denotes a conditional probability for a discrete variable or a conditional density for a continuous variable. The comparability-based criterion focuses on the causal effects of interest rather than an association measure. However, this criterion has to rest on the untestable conditions because Y_0 is not observed for the exposed, and $\text{pr}(Y_0 = y | X = 1, V)$ and $\text{pr}(Y_0 = y | X = 1)$ cannot be identified from observed data without further assumptions. For detailed discussions on the comparability-based criterion, readers are directed to Greenland & Robins (1986), Wickramaratne & Holford (1987), and Geng et al. (2001, 2002). Under the assumption $Y_0 \perp\!\!\!\perp X \,|\, V$, it can be shown that $V \not\perp\!\!\!\perp Y \,|\, X = 0$ and $V \not\perp\!\!\!\perp X$ is a necessary condition for V to be a confounder by using the comparability-based criterion, but it is not sufficient. Without requiring $Y_0 \perp\!\!\!\perp X | V$, Geng et al. (2002) showed the same necessary condition of a confounder in the sense that adjusting for a confounder can reduce confounding bias.

In contrast, the collapsibility-based criterion requires the assumption that the association measure used in the definition of collapsibility is free of confounding bias conditional on the observed covariate V. The criterion depends on what association measure or parameter is used and can be tested with observational data, but it does not involve any notion of causal effects. For example, the relative risk is collapsible over a certain covariate, but the odds ratio may not be. Discussions on the collapsibility-based criterion for various specific association measures are provided by Whittemore (1978), Geng (1992), Cox & Wermuth (2003), Ma et al. (2006), and Xie et al. (2008).

Greenland et al. (1999b), Geng et al. (2001), and Geng et al. (2002) provided comprehensive discussion on the relationships between the comparability-based, the collapsibility-based, and

Miettinen & Cook's criterion. Miettinen and Cook's criterion can be shown formally by using the comparability-based criterion, while the collapsibility-based criterion depends on what association measure or parameter is used. Both of the two criteria require the untestable assumption that $Y_0 \perp\!\!\!\perp X|V$, although under this assumption, their conditions for confounders are different.

Geng et al. (2001) proposed the concepts of occasional confounders and uniform nonconfounding of the coarse subpopulations. If we have $\text{pr}(Y_0 = y|X = 1, V \in \omega) \neq \text{pr}(Y = y|X = 0, V \in \omega)$ for some interval or coarse category ω of a covariate V, we say that V is an occasional confounder or that the population is not uniformly nonconfounding. For example, age is an occasional confounder if there is no confounding bias in the age groups of every 20 years but there is confounding bias in the age groups of every 10 years (see the example given in section 3.1 of Geng et al. 2002). It means that the population is not uniformly nonconfounding over any categorization of age. Under the assumption $Y_0 \perp\!\!\!\perp X|V$, Geng et al. (2001) showed that $V \not\perp\!\!\!\perp Y|X = 0$ and $V \not\perp\!\!\!\perp X$ is a necessary and sufficient condition for V to be an occasional confounder, but it is not a sufficient condition of a confounder. In the absence of occasional confounding, causal effects for the whole population, subpopulations, and coarse subpopulations are all identifiable, although these effects may differ (i.e., the causal effects may not be collapsible).

Another criterion for determining confounders is based on the causal diagrams (Greenland et al. 1999a). In a causal diagram, a variable set that blocks all of the backdoor paths from the treatment to the outcome is called a sufficient confounder set. Given a causal diagram, this criterion can exactly determine whether or not a variable is a confounder. In **Figure 1**, neither Z nor F is a confounder even if Z and F satisfy the necessary condition of confounders under the comparability-based criterion.

In many applications, however, it is difficult to have the knowledge of a complete causal diagram. Geng & Li (2002) and VanderWeele & Shpitser (2011) proposed criteria that do not require a complete causal diagram but require prior information about the potential confounders. Greenland et al. (1999a) and Wang et al. (2009) extended Miettinen & Cook's criterion to remove nonconfounders from a large sufficient confounder set V.

3. ADJUSTMENT FOR CONFOUNDERS

3.1. Adjustment for Observed Confounders

In observational studies, adjustment for observed confounders has rested on the treatment assignment ignorability assumption (Rubin 1978, Rosenbaum & Rubin 1983b): $Y_x \perp\!\!\!\perp X|V$, where V denotes the observed confounders. The assumption means that potential outcomes are independent of the treatment assignment mechanism conditional on the observed confounders. Further assuming that the propensity score $\text{pr}(X = 1|V)$ is bounded, i.e., $0 < \text{pr}(X = 1|V) < 1$, then the ACE can be identified by first evaluating the causal effect in each group separately and then averaging them:

$$\text{E}(Y_1 - Y_0) = \text{E}\{\text{E}(Y_1 - Y_0|V)\} = \text{E}\{\text{E}(Y|V, X = 1) - \text{E}(Y|V, X = 0)\}.$$

Figure 1

Neither Z nor F is a confounder based on the causal diagrams (Greenland et al. 1999a), although they satisfy the compatibility-based criterion.

The ignorability assumption is consistent with the backdoor criteria using the language of causal networks (Pearl 1995): The observed covariates V blocking all backdoor paths from X to Y should be conditioned to calculate the ACE.

The treatment assignment ignorability assumption provides the basis of adjustment for confounders in observational studies. Methods of adjustment for confounders can be classified into two categories. One is to balance the confounder distribution between the treated and control groups, which motivates stratification, matching, and inverse probability weighting methods. The other is to adjust the causes of the outcome, which motivates the regression estimation.

Stratification and matching are conducted by pairing treated and control units that are similar in terms of the observed confounders, to reduce imbalance of confounders, i.e., deviation of the confounder distribution between the treated and control groups. In early causal inference texts, stratification and matching methods are conducted by pairing units based on a single variable or weighting several variables (Cochran & Rubin 1973, Rubin 1973); however, it is impossible to do so with a large set of confounders. In this case, one approach is to model the relationship between the exposure and confounders. The propensity score reduces the dimensionality of matching to a single dimension and has become popular in causal inference (Rosenbaum & Rubin 1983b, Stuart 2010). Propensity score matching can be viewed as a nonparametric method of adjustment (Hahn 1998, Abadie & Imbens 2006). This approach can capture the treatment effect heterogeneity by separately estimating the effect in the treated and control groups, and it is easily implemented by software routines. However, it is difficult to estimate the variance of propensity score matching estimators (Abadie & Imbens 2006), and imbalance of the confounder distribution may increase when the treatment and confounders are weakly correlated (King & Nielsen 2016).

Inverse probability weighting (Horvitz & Thompson 1952, Robins et al. 1994) rests on a propensity score model $\pi(V;\alpha) = \mathrm{pr}(X = 1|V;\alpha)$ and evaluates the potential outcome mean by $\mathrm{E}(Y_1) = \mathrm{E}\{XY/\pi(V;\alpha)\}$ and $\mathrm{E}(Y_0) = \mathrm{E}\{(1-X)Y/\{1-\pi(V;\alpha)\}\}$. In contrast, regression adjustment rests on a regression model $\mathrm{E}(Y|X,V) = m(X,V;\gamma)$, and evaluates the potential outcome mean by $\mathrm{E}(Y_x) = \mathrm{E}\{m(x,V;\gamma)\}$. Regression and inverse probability weighting are perhaps the most common forms of data analysis in causal inference; they have the advantage of conceptual and computational simplicity and the ease of variance estimation. However, inefficiency and lack of weight stabilization are often encountered in implementation of inverse probability weighting, and regression adjustment involves interpolation and extrapolation and is more likely to lead to fishing expeditions or publication bias (Pocock et al. 2002, Rubin 2008)—that is, researchers may try different regression models and tend to report one resulting in a significant estimate of the causal effect.

All of these adjustment methods are subject to potential bias due to model misspecification. To enhance robustness, semiparametric models and flexible machine learning methods can be applied (Lee et al. 2010). Hybrid inference strategies of using both a propensity score model and a regression model, such as doubly robust estimation (Scharfstein et al. 1999) and g-estimation (Robins et al. 1992), are used to improve the efficiency and robustness against model misspecification, although they may have larger bias when both the propensity score and the regression models are incorrect (Kang & Schafer 2007).

Apart from model misspecification, lack of common support of the confounder distribution between the treated and control groups diminishes the credibility of all confounder adjustment methods because for the units with propensity scores close to zero or unity, the causal effect is not identified. In this case, adjustment for confounders involves either inconsistent extrapolation or extreme weighting that results in poor estimation. Trimming, or discarding problematic units that have extreme propensity scores, has been suggested as a way to deal with this problem (Heckman

et al. 1998, Crump et al. 2009). However, this approach is sensitive to the level of trimming, and it changes the target population and the interpretation of the estimand. Petersen et al. (2012), Hill & Su (2013), and Fogarty et al. (2016) discussed the strategies for detecting violations of common confounder support and for adjusting the target population of estimation.

3.2. Adjustment for Unobserved Confounders

When there exist unobserved confounders, selection bias, or measurement error, the treatment assignment ignorability does not hold and $Y_x \not\perp\!\!\!\perp X \mid V$, and thus, the adjusting methods for observed confounders cannot effectively remove the confounding bias. Alternatively, it is more reasonable to assume the latent ignorability assumption (Frangakis & Rubin 1999):

$$Y_x \perp\!\!\!\perp X|(U,V), \quad 0 < \text{pr}(X|U,V) < 1,$$

where U and V denote the unobserved and observed confounders, respectively. For notation convenience, we suppress the observed confounders V hereafter.

Under latent ignorability, the potential outcome mean can be evaluated by

$$\text{E}(Y_x) = \text{E}\{\text{E}(Y|U,x)\}.$$

The crucial problem of implementing this formula is that U is not observed, and the conditional mean $\text{E}(Y|U,x)$, the probability density or mass function $\text{pr}(U)$, and the propensity score $\text{pr}(X|U)$ cannot be determined from the observed data. This results in the challenging problem of identification: The causal effect is not uniquely determined, even if the sample size is arbitrarily large. Auxiliary variables are indispensable to identify causal effects in the presence of unobserved confounders. Below, we review two approaches that exploit auxiliary variables to eliminate the confounding bias due to unobserved confounders: the influential approach of using an instrumental variable and the negative control approach that has recently captured the interest of researchers.

3.2.1. Adjustment with an instrumental variable. The instrumental variable approach, originating in the econometrics literature in the 1920s, is a popular method to mitigate the problem due to unobserved confounders or endogeneity in observational studies. This approach exploits an auxiliary covariate Z that satisfies the following:

(*i*) no direct effect on the outcome, $Z \perp\!\!\!\perp Y|(X,U)$ (exclusion restriction)
(*ii*) independent of the unobserved confounders, $Z \perp\!\!\!\perp U$ (independence)
(*iii*) associated with the exposure, $Z \not\perp\!\!\!\perp X$ (relevance)

If the instrumental variable is an exposure or treatment assignment that occurs before the primary exposure X, i.e., a causal instrumental variable, then we can define the potential exposure X_z that will arise if Z is set to z by external intervention, and the potential outcome Y_{zx} that would be observed if X is set to x and Z is set to z. The following conditions provide an alternative definition of an instrumental variable:

(*a*) exclusion restriction: $Y_{zx} = Y_x$
(*b*) independence: $(X_z, Y_{zx}) \perp\!\!\!\perp Z$
(*c*) relevance: $E(X_z)$ is a nontrivial function of z

For instance, noncompliance occurs in some randomized clinical trials. The treatment actually received, X, is affected by but not necessarily identical to the assigned treatment Z, and the outcome Y is only affected by the received treatment. In this case, the treatment assignment Z is an instrumental variable, which is independent of the latent confounder U and is associated with the treatment actually received, X.

Under the three core conditions, the ACE is only partially identified, that is, only certain upper and lower bounds can be derived (Manski 1990, Balke & Pearl 1997). Unfortunately, such bounds are often wide and include the null value, which cannot produce informative results. Identification of causal effects using an instrumental variable requires extra model assumptions. A commonly used assumption is effect homogeneity (Hernán & Robins 2018), which is encoded in structural equation models (Wright 1928, Goldberger 1972) or structural mean models (Robins 1994). For example, the linear regression model $\mathrm{E}(Y \mid X, U) = \gamma_1 X + U$ encodes a constant causal effect that is γ_1 for all individuals. The conventional instrumental variable estimator under this model is $\gamma_1^{\mathrm{iv}} = \widehat{\sigma}_{zy}/\widehat{\sigma}_{xz}$, where $\widehat{\sigma}_{zy}$ is the sample covariance of Z and Y. Under the instrumental variable conditions *i–iii*, the bias of γ_1^{iv} converges to zero under a large sample size because $Z \perp\!\!\!\perp U$. That is the key reason instrumental variable estimation can eliminate the confounding bias.

An alternative assumption is that the effect of Z on X is monotone, i.e., $X_{z=1} \geq X_{z=0}$, which means that there is no one who does the opposite of his assignment. Under conditions *a–c* and the monotonicity assumption, the complier ACE $\mathrm{E}(Y_1 - Y_0 | X_1 = 1, X_0 = 0)$ is identified (Imbens & Angrist 1994, Angrist et al. 1996). Recent surveys of the instrumental variable approach and comparison of the homogeneity and monotonicity assumptions include those of Hernán & Robins (2006), Clarke & Windmeijer (2012), and Hernán & Robins (2018).

Recent research on the instrumental variable approach has centered around validity checking and violation detection of the core conditions and corresponding mitigation methods. Balke & Pearl (1997) and L. Wang et al. (2017) proposed falsification tests of the instrumental variable assumptions. Bound et al. (1995), Stock et al. (2002), and many others discussed the weak instrument problem (i.e., Z and X are weakly correlated). Manski & Pepper (2000) and Small (2007) proposed bounding and sensitivity analysis methods under violations of the exclusion restriction assumption. Bowden et al. (2015) and Kolesár et al. (2015) discussed estimation methods using multiple or many invalid instrumental variables in economics and Mendelian randomization studies. Lin et al. (2015) and Kang et al. (2016) considered variable selection and estimation with high-dimensional instrumental variables.

3.2.2. Adjustment with negative controls. Negative control variables are those not causally associated with the primary treatment nor outcome, but correlated with the unobserved confounders. The tradition of using negative controls in causal inference dates back to the notion of specificity initiated by Hill (1965) and Yerushalmy & Palmer (1959). As Hill (1965) advocated, if one observes that the exposure has an effect only on the primary outcome but not on other ones, then the credibility of causation is increased; Weiss (2002) emphasized that in order to apply Hill's specificity criterion, one needs prior knowledge that only the primary outcome ought to be causally affected by the exposure. Rosenbaum (1989), Lipsitch et al. (2010), and Flanders et al. (2011) described guidelines for using negative control variables or known effects to detect confounding in observational studies. Schuemie et al. (2014) discussed using negative controls for p-value calibration in medical studies. Empirical negative control studies include those of Trichopoulos et al. (1983), Davey Smith (2008), and Flanders et al. (2017). But confounding detection and robustness checks do not make full use of the negative control variables and cannot give definitive conclusions about causal associations or identify the causal effect.

Miao & Tchetgen Tchetgen (2017) and Miao et al. (2018) formally established the framework of using negative controls for unobserved confounder adjustment. Let U denote the unobserved confounder such that the latent ignorability holds, i.e., $Y_x \perp\!\!\!\perp X \mid U$ and $0 < \mathrm{pr}(X = 1 \mid U) < 1$. Negative control variables are classified into two classes: negative control outcome W,

$$W \perp\!\!\!\perp X | U, \quad W \not\perp\!\!\!\perp U,$$

and negative control exposure Z,

$$Z \perp\!\!\!\perp Y \mid (U, X), \quad Z \perp\!\!\!\perp W \mid (U, X).$$

Figure 2 provides a causal network to depict the conditional independencies and to illustrate negative controls.

Because X does not directly affect W, the association between (X, W) is determined by their associations through U. A nonzero association between (X, W) indicates the presence of unmeasured confounders, and based on this property, previous authors used negative control outcomes for confounding detection or robustness checks.

Given the above conditions, negative controls can be used as proxies of the unobserved confounder, and Gagnon-Bartsch & Speed (2012), Flanders et al. (2017), and J. Wang et al. (2017) explored confounding bias correction under their respective model assumptions. Miao & Tchetgen Tchetgen (2017) and Miao et al. (2018) described general identification conditions and confounding adjustment methods with negative controls: For any (y, z, x), they proposed to solve

$$\mathrm{pr}(Y = y \mid Z = z, X = x) = \int_{-\infty}^{+\infty} h(w, x, y)\mathrm{pr}(W = w \mid Z = z, X = x)\mathrm{d}w$$

for $h(w, x, y)$, with $\mathrm{pr}(Y = y \mid Z = z, X = x)$ and $\mathrm{pr}(W = w \mid Z = z, X = x)$ estimated from observed data. Then, under certain rank or completeness conditions, the potential outcome distribution is identified by

$$\mathrm{pr}(Y_x = y) = E\{h(W, x, y)\},$$

and the ACE is identified.

The negative control approach can be used to eliminate the bias of the instrumental variable estimator. In the instrumental variable design, independence between the instrumental variable and the confounder may be violated in practice, which results in a biased estimator. However, by incorporating a negative control outcome and treating the instrumental variable as a negative control exposure, one can apply the negative control adjustment to identify the causal effect, and the bias due to the invalid instrumental variable is removed.

4. SURROGATE PARADOX AND CRITERIA FOR SURROGATES

4.1. Surrogates and Their Criteria

In clinical trials, surrogates are often used to assess treatment effects on unobserved endpoints when measurement of the endpoints is expensive or infeasible. There have been a number of papers questioning the validity of surrogates (Fleming & Demets 1996, Baker 2006, Manns et al. 2006, Alonso & Molenbergh 2008). They pointed out that in many real clinical trials, the application of surrogates has falsely evaluated treatment effects on endpoints, such as using CD4 counts as a

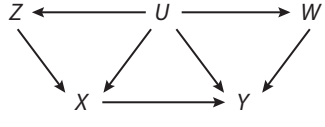

Figure 2

A graph model where Z and W denote a pair of negative control exposure and outcome, respectively.

surrogate for survival time in clinical trials of AIDS and bone mass as a surrogate for fracture in osteoporosis studies.

There are many approaches for using surrogates to quantitatively evaluate the treatment effects on endpoints (Buyse & Molenberghs 1998, Burzykowski et al. 2005). The quantitative approaches require that the measures of associations in the previous validation studies can be transferred quantitatively to the future studies. For example, the proportion of the treatment effect proposed by Freedman et al. (1992) is used to assess the surrogate validation based on associations among variables obtained in previous validation studies. Li et al. (2010) proposed a Bayesian approach for assessing principal surrogates.

If the endpoint of interest is never observed in clinical trials on the treatment, we need prior knowledge about relationships among treatment, endpoint, and surrogate. Based on the prior knowledge, we use the treatment effect on the observed surrogate to qualitatively predict the sign of the treatment effect on the unobserved endpoint. There are several criteria for qualifying surrogates. The most intuitive criterion may be that there is a strong association between the endpoint and the surrogate. However, the strong association alone is not sufficient for using surrogates to evaluate the treatment effect on the endpoint. For example, there is a strong association between the shoe size and the reading ability for primary school children. However, changing into a larger pair of shoes does not increase a child's reading ability.

Prentice (1989) proposed a criterion for surrogates that requires both a strong association and the condition that the endpoint Y is independent of a treatment X conditionally on a surrogate S, denoted as $Y \perp\!\!\!\perp X \mid S$. The surrogates satisfying Prentice's criterion are called the statistical surrogates. Conditional independence means that the surrogate S can break the association between the treatment X and the endpoint Y, and thus, the association between X and Y can be represented by the association between X and S if S and Y have a strong association. Under his criterion, Prentice (1989) showed that $X \perp\!\!\!\perp S$ implies $X \perp\!\!\!\perp Y$, although the reverse may not be true. Thus, for a randomized trial of treatment X, Prentice's criterion can ensure that a null treatment effect on the surrogate S (i.e., $\text{pr}(S_1 = s) = \text{pr}(S_0 = s)$, which is equivalent to $\text{pr}(S = s \mid X = 1) = \text{pr}(S = s \mid X = 0)$) implies a null treatment effect on the endpoint Y (i.e., $\text{pr}(Y_1 = y) = \text{pr}(Y_0 = y)$, which is equivalent to $\text{pr}(Y = y \mid X = 1) = \text{pr}(Y = y \mid X = 0)$).

Frangakis & Rubin (2002) presented the idea that a surrogate should possess the property of causal necessity: A treatment has a causal effect on endpoint Y only if the treatment has an ICE on the surrogate S. They gave a numerical example in which the statistical surrogate S does not possess the property of causal necessity: For individuals in the principal stratum defined by $S_{X=1} = S_{X=0}$, there is no individual treatment effect on the statistical surrogate S, but there may be individual, distributional, and average treatment effects on the endpoint Y. To make a surrogate possess the property of causal necessity, Frangakis & Rubin (2002) defined that S is a principal surrogate if for all fixed s, the comparison of elements between the ordered sets

$$\{Y_{X=1}(i) : \text{for every individual } i \text{ such that } S_{X=1}(i) = S_{X=0}(i) = s\}$$
$$\{Y_{X=0}(i) : \text{for every individual } i \text{ such that } S_{X=1}(i) = S_{X=0}(i) = s\}$$

results in equality. The potential outcomes of $Y_{X=1}(i)$ and $Y_{X=0}(i)$ are compared pairwise in the order of individuals in the ordered sets.

Lauritzen (2004) proposed the strong surrogate criterion that the surrogate S breaks the causal path from treatment X to endpoint Y in a causal diagram. A strong surrogate also satisfies the causal necessity. If treatment is binary, then a principal surrogate and a strong surrogate are equivalent. Gilbert & Hudgens (2008) proposed that any reasonable surrogate should have both the average causal necessity and the average causal sufficiency. Joffe & Greene (2009) summarized related statistical approaches and discussed the relationships among these approaches.

4.2. Surrogate Paradox

For statistical surrogates, principal surrogates, and strong surrogates, Chen et al. (2007) showed that a treatment X may have a positive causal effect on a surrogate S, and the surrogate may have a positive causal effect on an endpoint Y, but the treatment X may have a negative effect on the endpoint Y. This phenomenon is called the surrogate paradox. The surrogate paradox means that the sign of the treatment effect on the endpoint cannot be predicted by the sign of treatment effect on the surrogate and the sign of the causal effect of the surrogate on the endpoint.

The surrogate paradox presents a key issue of the surrogate approach. A surrogate S is an intermediate variable in the causal path from X to Y, and variable X can be seen as an instrumental variable. A more proper name for the paradox might be "the intermediate variable paradox" to reflect a wider generality. The same paradox applies to other situations. For example, the paradox can be called the instrumental paradox in the use of instrumental variable. This paradox also points out the issue of the transitivity of causal effects on a causal path.

Moore (1995) gave a real-life example of the surrogate paradox. Doctors know that having an irregular heartbeat is a risk factor for sudden death, and they presumed that correcting irregular heartbeats would prevent sudden death. Thus, with correction of heartbeat as a surrogate, several drugs (encainide, flecainide, and moricizine) were approved by FDA. However, the Cardiac Arrhythmia Suppression Trial (CAST 1989) showed that these drugs did not improve survival times and instead increased mortality.

Below, we give a numerical example to illustrate the surrogate paradox. Let X denote the treatment, 1 for treated and 0 for control. Let S denote the correction of an irregular heartbeat, 1 for corrected and 0 for uncorrected, and S_x the potential outcome of correction under treatment $X = x$. The endpoint Y is the survival time, and Y_{sx} is the potential survival time under $X = x$ and $S = s$. We assume that the treatment effect on survival time is completely driven by the correction of an irregular heartbeat, that is, $Y_{sx} = Y_{sx'} = Y_s$. Thus, we have that $S_1 = S_0 = s$ implies $Y_{s0} = Y_{s1}$, and the heartbeat status S is a principal surrogate and also a strong surrogate. Further assume that correction of an irregular heartbeat has a positive individual effect on the endpoint, i.e., $Y_{S=1}(i) > Y_{S=0}(i)$ for every patient i. These two assumptions are very stringent, but we can still show the occurrence of the surrogate paradox using the artificial population with 100 patients with an irregular heartbeat in **Table 1**. For binary S_0 and S_1, the 100 patients are stratified into four principal strata (1 to 4) as shown in columns 3 and 4: never corrected ($S_0 = 0, S_1 = 0$), where the heartbeat status S will not be corrected regardless of the treatment X; effectiveness ($S_0 = 0, S_1 = 1$), where S will be corrected if and only if the patient is assigned to the treated group, $X = 1$; defiers ($S_0 = 1, S_1 = 0$), where S will be corrected if and only if the patient is assigned to the control group, $X = 0$; and always corrected ($S_0 = 1, S_1 = 1$), where S will always be corrected regardless of X. The number of patients in every stratum is given in column 2. For

Table 1 Illustration of the surrogate paradox for 100 patients with irregular heartbeat

Stratum	N of patients	$S_{X=0}$	$S_{X=1}$	$Y_{S=0}$	$Y_{S=1}$	$Y_{X=0}$	$Y_{X=1}$
1	20	0	0	3	5	3	3
2	40	0	1	6	7	6	7
3	20	1	0	5	8	8	5
4	20	1	1	9	10	10	10

S denotes the correction of an irregular heartbeat, with 1 for corrected and 0 for uncorrected. X is the treatment, with 1 for treated and 0 for control. Y is the survival time in years.

simplicity, we assume that all patients in the same principal stratum have the same potential survival time. Columns 5 and 6 give the potential survival times for patients in all principal strata, although some of them are prior counterfactual. For example, $Y_{S=1}$ in stratum 1 is prior counterfactual since $S = 1$ cannot be obtained by assigning either treatment $X = 0$ or 1 unless the external intervention $S = 1$. From columns 3 to 6, we can obtain the potential survival times $(Y_{X=0}, Y_{X=1})$. For example, if a patient in stratum 2 were treated, then the patient would have $S_{X=1} = 1$ and thus $Y_{X=1} = Y_{S_{X=1},X=1} = Y_{S_{X=1}} = Y_{S=1} = 7$. For the population of the 100 patients given in **Table 1**, we have that the ACE of treatment on the correction of irregular heartbeats is positive,

$$\text{ACE}_{X\to S} = \frac{40+20}{100} - \frac{20+20}{100} = \frac{20}{100} > 0,$$

but we obtain a negative treatment effect on the survival time

$$\text{ACE}_{X\to Y} = \frac{3\cdot 20 + 7\cdot 40 + 5\cdot 20 + 10\cdot 20}{100} - \frac{3\cdot 20 + 6\cdot 40 + 8\cdot 20 + 10\cdot 20}{100} = -\frac{20}{100} < 0.$$

In the above example, the surrogate paradox occurs since there are defiers ($S_{X=0} = 1, S_{X=1} = 0$), that is, for 20 patients the treatment did not correct the irregular heartbeat. Geng (2015) gave other examples of the surrogate paradox for the case without defiers and for the statistical surrogate.

4.3. Criteria for Consistent Surrogates

To avoid the surrogate paradox, Chen et al. (2007), Ju & Geng (2010), Wu et al. (2011), and VanderWeele (2013) presented the criteria for consistent surrogates, which require that the sign or direction of treatment effect on a surrogate can be used to predict the sign or direction of treatment effect on the unobserved endpoint. Assume that treatment X is randomized. Since the surrogate S may not be randomized, there may be an unobserved confounder or a confounder set U that affects both the surrogate S and the endpoint Y. Let Y_x and S_x denote the potential outcomes of the endpoint and the surrogate under a treatment $X = x$, respectively. For a continuous surrogate S, we denote the distributional causal effect (DCE) of X on the endpoint S as $\text{DCE}_{X\to(S>s)} = \text{pr}(S_1 > s) - \text{pr}(S_0 > s)$. For example, the DCE is used to assess the causal effect of smoking (X) on hypertension (blood pressure larger than a threshold s). We say that X has a nonnegative (null) DCE on S if $\text{DCE}_{X\to(S>s)} \geq (=) 0$ for all s. If, in addition, there exists a threshold s such that $\text{DCE}_{X\to(S>s)} > 0$, then we say that X has a positive DCE on S.

We now introduce two criteria to avoid the surrogate paradox, one based on the prior knowledge of causation and the other based on the prior knowledge of association. First, we introduce the criterion for the consistent surrogates based on the causation knowledge. If

1. the ACE of S on Y conditional on $U = u$ is nonnegative for all u, and
2. the sign of DCE of X on S conditional on $U = u$ does not change with u,

then a nonnegative DCE of X on S implies a nonnegative ACE of X on Y, and a null DCE of X on S implies a null ACE of X on Y. Furthermore, if

3. conditional on some $U = u$, both the ACE of S on Y is positive and the DCE of X on S has the strict inequality sign ("> 0" or "< 0"),

then a positive DCE of X on S implies a positive ACE of X on Y.

Condition 1 means that S is a risk factor for the endpoint Y conditional on any u. For example, we would require that irregular heartbeat is a risk factor for sudden death for all values of the unobserved confounder. Condition 2 means that the conditional DCE of treatment X on the surrogate S given $U = u$ has the same sign for all u, although we may not know whether it is

positive or negative. Condition 3 means the strict signs of inequality (>0 or <0) of conditional causal effects of X on S and of S on Y given some $U = u$ simultaneously hold so that the causal effect of X on Y conditional on some u has a strict positive or negative sign, and its probability is not zero. Ju & Geng (2010) showed how several commonly used models of Y and of S (e.g., generalized linear models, proportional hazards models, and some of their extended models) including a latent variable U can avoid the surrogate paradox. In these models, a positive causal effect of X on S implies that the coefficient of X in the models of S is positive, and then both conditions 2 and 3 are satisfied; condition 1 is satisfied by a positive coefficient of S in the models of Y, and thus, the effect sign of X on S can predict the effect sign of X on Y.

Let S be a continuous strong surrogate that breaks the path from treatment X to endpoint Y so that X is an instrumental variable, and let $\mathrm{ACE}_{S\to Y}(s,s') = \mathrm{E}(Y_s) - \mathrm{E}(Y_{s'})$. Chen et al. (2007) proved the equation of ACEs for $s > s'$,

$$\mathrm{ACE}_{X\to Y} = \frac{\mathrm{ACE}_{X\to S}\mathrm{ACE}_{S\to Y}(s,s')}{s-s'},$$

under either of the following two models.

- Model I: a nonparametric model for $\mathrm{pr}(s|x,u)$ and a linear model for Y,

$$\mathrm{E}(Y|S=s,U=u) = \alpha_1 s + \gamma_1(u)$$

- Model II: semi-parametric models for S and Y,

$$\mathrm{E}(S|X=x,U=u) = \beta_2(x) + \gamma_2(u),$$
$$\mathrm{E}(Y|S=s,U=u) = s\beta_1(u) + \gamma_1(u)$$

Here, α_i, β_i, and γ_i denote parameters, and $\alpha_i()$, $\beta_i()$, and $\gamma_i()$ denote unknown functions. From this equation of ACEs, we can obtain an instrumental variable estimation equation of $\mathrm{ACE}_{S\to Y}(s,s')$ for models I and II with an unobserved confounder U between S and Y.

Since U is never observed, these conditions and models cannot be tested from observed data even if Y is observed in a validation study, and thus we need the prior knowledge about the causation relationships among these variables. The causation-based criterion requires that all paths from treatment X to the endpoint Y are broken by a single intermediate variable S, which means no direct effect from X to Y, that is, there are no other paths from X to Y. This requirement rather than a conditional independency cannot be checked by data. The advantage of the causation-based criterion for the consistent surrogates is that the prior knowledge of the causal mechanisms may be obtained from experts, but the disadvantages are the empirical untestability and the requirement of no direct effect of the treatment on the endpoint.

Next, we introduce the criteria for consistent surrogates based on the prior knowledge of the association relationships among treatment, surrogate, and endpoint. If

1. $\mathrm{E}(Y|s,X=0) - \mathrm{E}(Y|s',X=0) \geq 0$ for all $s > s'$ or
   $\mathrm{E}(Y|s,X=1) - \mathrm{E}(Y|s',X=1) \geq 0$ for all $s > s'$, and
2. $\mathrm{E}(Y|s,X=1) - \mathrm{E}(Y|s,X=0) \geq 0$ for all s,

then a nonnegative DCE of treatment X on S can predict a nonnegative ACE of treatment X on Y.

Unlike the conditions based on causation, these conditions based on association are testable if the endpoint Y is observed in a validation study, and they allow for the direct effect of the treatment on the endpoint. Condition 1 requires that the monotonicity property of the expectation of Y in s holds for only one treatment group, not necessarily for both groups. Particularly for the control group of $X = 0$, the control treatment (such as a placebo) may have been used in the

previous trials, and thus the prior knowledge about condition 1 for $X = 0$ may be obtained from the previous trials. Condition 2 describes the monotonicity property in x that the treatment X nonnegatively associates with the endpoint conditional on the surrogate. When Prentice's conditional independence $Y \perp\!\!\!\perp X \mid S$ holds, condition 2 is satisfied. This criterion shows that to avoid the surrogate paradox, we need to add one more condition to Prentice's criterion: The conditional expectation of the endpoint in the control group is a monotonic function of the surrogate S. For example, irregular heartbeat is a risk factor for sudden death in the placebo group, but it is not required that irregular heartbeat must also be a risk factor for sudden death in the treated group. Furthermore, under the assumptions of generalized linear models and the exponential distribution family, the treatment effects on the surrogate and the endpoint have the same signs (positive, negative, or null) (Ju & Geng 2010, Wu et al. 2011).

All of the above criteria are only for a single surrogate. In many applications, however, a treatment may affect the endpoint through several paths, and thus a single surrogate cannot break these paths. For example, a drug may reduce a death due to AIDS through two paths, by decreasing HIV-1 RNA and by increasing CD4 count. In this case, a single surrogate may not satisfy any criteria of the statistical, principal, strong, and consistent surrogates, and both HIV-1 RNA and CD4 count should be used as multiple surrogates for death due to AIDS. Joffe (2013) suggested that it is meaningful to generalize the criteria for a single surrogate to multiple surrogates. Luo et al. (2018) proposed a criterion for multiple surrogates $\mathbf{S} = (S_1, \ldots, S_p)$ based on stochastic orders of random vectors. We depict a causal network with $p = 2$ surrogates S_1 and S_2 in **Figure 3**. The double-headed arrow between S_1 and S_2 means that they are correlated. Let $\mathbf{S}(x)$ denote the potential outcome of multiple surrogates $\mathbf{S}$ under treatment x. Let $f(\mathbf{s}, x) = \mathrm{E}(Y \mid \mathbf{s}, x)$, and say that $f(\mathbf{s}, x)$ increases in a vector $\mathbf{s} = (s_1, \ldots, s_p)$ if it increases in every element s_i for $i = 1, \ldots, p$. Luo et al. (2018) gave the following criterion for multiple surrogates to avoid the surrogate paradox. Suppose that we have the following knowledge about the conditional expectation $f(\mathbf{s}, x)$:

1. $f(\mathbf{s}, 1)$ or $f(\mathbf{s}, 0)$ is a nonconstant increasing function of $\mathbf{s}$, and
2. $f(\mathbf{s}, 1) \geq f(\mathbf{s}, 0)$ for all $\mathbf{s}$.

Under this supposition, Luo et al. (2018) showed that if $\mathbf{S}(1)$ is larger than $\mathbf{S}(0)$ in the stochastic order, then the causal effect of treatment X on the endpoint Y is positive. Thus, we can use the signs of treatment effects on multiple surrogates $\mathbf{S}$ to predict the sign of treatment effect on the unobserved endpoint Y. All of these conditions in the multiple-surrogate criterion can be tested if there is a validation trial in which the endpoint is observed, and some of these conditions can also be tested if the endpoint has been observed in a previous trial with the same placebo. Luo et al. (2018) also gave the sufficient conditions for the sign equivalence of treatment effects on the surrogates and on the endpoint under the conditional independence of treatment and the endpoint given multiple surrogates. Furthermore, they illustrated that the multiple-surrogate criterion can be satisfied by many commonly used models.

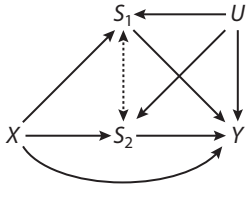

Figure 3

Causal network with treatment X, two surrogates S_1 and S_2, endpoint Y, and latent confounder U.

5. CAUSAL NETWORKS AND STRUCTURE LEARNING

5.1. Causal Networks

Causal effect evaluation, mentioned in previous sections, focuses on the causal effects of a given treatment on a given outcome. In many scientific studies, however, researchers may be interested in the associations, interactions, and causal relationships among multiple variables. To formulate the associations and interactions among multiple discrete variables, log-linear models were pioneered by Birch (1963), Goodman (1970), Haberman (1974), Bishop et al. (1975), and Fienberg (1980), which are widely used for discrete multivariate analysis. Darroch et al. (1980) presented graphical models that depict conditional independencies via the absences of edges between variables in undirected graphs. Fienberg (2011) reviewed the developments of log-linear models and graphical models.

To formulate causal relationships and data generating mechanisms, Pearl (1988, 2009) presented causal networks depicted by directed acyclic graphs (DAGs). We let $\{X_1, \ldots, X_p\}$ denote the vertex set of a DAG of interest, a directed edge $X_i \to X_j$ denotes that X_i is a parent or cause of X_j and that X_j is a child or effect of X_i, and $\text{pa}(X_i)$ denotes the parent set of X_i. The data generating mechanism can be represented by $X_i = f_i(\text{pa}(X_i), \varepsilon_i)$ for every i, where ε_i is an independent random residual variable that does not affect any other variable X_j for any $j \neq i$. Then, the joint distribution of variables can be factorized into

$$\text{pr}(X_1, \ldots, X_p) = \prod_{i=1}^{p} \text{pr}(X_i \,|\, \text{pa}(X_i)).$$

We make the faithfulness assumption, which is used for structure learning of DAGs (Pearl 2009). Under the faithfulness assumption, a DAG can encode all conditional independencies of a distribution. A Markov equivalence class consists of all DAGs that encode the same set of conditional independencies. A Markov equivalence class can be depicted by an essential graph that has the same skeleton as all DAGs in the class and whose edges are directed if and only if the edges have the same orientation in all DAGs in the class (Verma & Pearl 1990).

5.2. Active Learning of Causal Networks

From observed data, we cannot distinguish the Markov equivalent DAGs from each other, but we can potentially find an essential graph depicting the Markov equivalence class (Andersson et al. 1997). In order to discover more causal relationships among variables, we need additional data from intervention experiments to orient the directions of edges in an essential graph and to reduce the number of candidate DAGs in the class. Cooper & Yoo (1999) presented a method of causal discovery from a mixture of experimental and observational data. Tian & Pearl (2001) proposed a method of discovering causal structures based on dynamic environment. Tong & Koller (2001) and Murphy (2001) presented active learning of network structures using a Bayesian framework. To determine uniquely the underlying causal network in a Markov equivalence class learned from observational data, He & Geng (2008) proposed two optimal designs of intervention experiments for the active learning approach: the batch-intervention design and the sequential-intervention design, to minimize the number of variables to be manipulated. Hauser & Bühlmann (2012) proposed the active learning approach for learning an interventional Markov equivalence class of DAGs from multiple interventions.

5.3. Local Structure Learning Around a Given Target Variable

In observational studies, researchers are often interested in discovering the causes and effects of a given variable rather than discovering a global causal network. For example, researchers want to

find the causes of lung cancer, the effects of smoking, and the causes and effects of hypertension. Guyon et al. (2008) organized a causation and prediction challenge focusing on predicting the causal effect of external interventions on a given variable. There are some algorithms that aim to find local causal relationships around a given variable X_i. The algorithms proposed by Peña et al. (2007) and Aliferis et al. (2010) find the parents and the children (PC) and the Markov blanket (MB) of X_i [denoted as MB(X_i)] so that the given variable is conditionally independent of other variables given MB(X_i). But these algorithms do not distinguish the parents (causes) from the children (effects) of the given variable.

The MB fan search algorithm (Ramsey 2006), the PCX algorithm (Bai et al. 2004), and the tabu search-enhanced MB (TS/MB) algorithm (Bai et al. 2008) learn a graphical MB model of a given variable X_i for classification. These algorithms can find all the edges connecting X_i and can orient some but not all edge directions.

Xie et al. (2006) and Xie & Geng (2008) presented the decomposition learning approach. Suppose the vertex set of a DAG of interest can be represented by the union of three disjoint subsets A, B, and C, such that $A \perp\!\!\!\perp B \mid C$. For $X_1 \in A$ and $X_2 \in A \cup C$, they are conditionally independent given a subset of $A \cup B \cup C$ if and only if they are conditionally independent given a subset of $A \cup C$. When both $X_1, X_2 \in C$, they are conditionally independent given a subset of $A \cup B \cup C$ if and only if they are conditionally independent given a subset of $A \cup C$ or given a subset of $B \cup C$. According to these results, Xie et al. (2006) proposed the decomposition learning algorithm, which decomposes the problem of searching conditional variable sets in the full set of all variables into the problems of searching them in many smaller variable sets. Xie & Geng (2008) presented the recursive algorithm, which recursively decomposes the problem of learning a large DAG into the problems of learning two smaller DAGs.

For the case that the set A contains a single variable X_i and C is MB(X_i) such that $X_i \perp\!\!\!\perp$ (other variables)|MB(X_i), then the above results can be used to learn the skeleton edges connecting X_i. Wang et al. (2014) proposed the MB-by-MB algorithm to find a local network centering around X_i. The MB-by-MB algorithm first finds MB(X_i) and constructs a local structure of MB(X_i), and then it sequentially finds MB(X_j) of every neighbor X_j connecting X_i and simultaneously constructs local structures of MB(X_j). This sequential process of finding an MB and then constructing a local structure of the MB along the paths starting from X_i is repeated until the causes and effects of X_i have been determined or the local undirected graph around X_i is surrounded by the directed edges. For the former case, it finds the direct causes and direct effects of X_i. For the latter case, it obtains a local structure with some undirected edges connecting X_i that cannot be oriented even if it learns the global essential graph.

In practice, to discover the causes of a given outcome or the effects of a given exposure, we need to design an observational study before collecting data. The MB-by-MB approach offers an idea for designing a sequential observational study. Let SuperMB(X_j) denote a superset that contains X_j and MB(X_j). Suppose that we know a sufficiently large set that contains MB(X_j) [i.e., SuperMB(X_j)]. In a sequential observational study, given the target variable X_i of interest, we first observe SuperMB(X_i) and learn a local structure over SuperMB(X_i); next we observe SuperMB(X_j) for a neighbor X_j of X_i in the local structure and learn a local structure over SuperMB(X_j). Repeat this MB-by-MB process of observing and learning until the causes and the effects of X_i are identified or the undirected graph around X_i is surrounded by directed edges. For the latter case, we may need some interventions to determine the directions of these undirected edges connecting X_i.

6. DISCUSSION

In this article, we reviewed the two frameworks for causal inference: the potential outcome framework and causal networks. Both of them are used to define and to evaluate causal effects. The

former focuses on inference about the effect of a given treatment or exposure, whereas the latter can be used to discover causal relationships among variables. Both the Yule-Simpson paradox and the surrogate paradox are induced due to latent confounders. The criteria for confounders are useful to identify the confounders to be adjusted for during data analysis. In addition, they can also be used in prospective study designs to determine what variables should be recorded. There is a debate about whether or not adjustment for a nonconfounder can improve the efficiency of causal inference. Mantel & Haenszel (1959), Gail (1986), Mantel (1989), Wickramaratne & Holford (1989), and Robinson & Jewell (1991) showed that adjustment for a nonconfounder in a linear regression model results in improved precision, whereas such adjustment in a logistic regression model results in a loss of precision. Breslow & Day (1980) also addressed how stratification by nonconfounders can increase the variability of the estimates of relative risk without eliminating any bias. Wang et al. (2007) compared the estimates of the counterfactual proportion $\text{pr}(Y_0|X = 1)$ with and without adjusting for a nonconfounder, and they showed that the variance of the proportion estimate is increased by adjusting for a nonconfounder when the sample size of the unexposed group is larger than the sample size of the exposed group.

In addition to adjustment and exact identification, sensitivity analysis (Cornfield et al. 1959, Rosenbaum & Rubin 1983a, Rosenbaum 2002) is a powerful tool to assess the impact of confounding and untestable assumptions on causal conclusions. Although plenty of methodological research and empirical applications exist for confounding adjustment and sensitivity analysis, extensions of such methods and novel approaches are necessary to accommodate complex data and integrate summary data in modern causal inference, for instance, two-sample instrumental variable estimation (Inoue & Solon 2010) and meta-analysis (DerSimonian & Laird 1986, Brockwell & Gordon 2001).

In our review of the surrogate paradox and the criteria for surrogates, we only discussed the qualitative evaluation of treatment effects, but in practice the quantitative evaluation may be more important. The quantitative approaches require quantitatively transferring the measures of associations from previous validation studies to the future studies (Freedman et al. 1992, Li et al. 2010). Both the qualitative and quantitative approaches have their respective merits and could therefore be used in parallel in real applications.

The causation-based criteria for surrogates require prior knowledge on the causal mechanism and an untestable assumption that the surrogate breaks all paths from the treatment to the endpoint. The association-based criteria rest on prior knowledge on association relationships between the treatment, surrogate, and endpoint, but not on the aforementioned assumption, and thus they may not satisfy the causal necessity. If the endpoint is observed in a validation study, the association-based criteria can be checked by observed data. But association knowledge is not always translated directly from causation knowledge. With certain models for associations and causations (e.g., the generalized linear model, proportional hazards models, and some of their extended models), the surrogate paradox may be avoided.

It is more challenging to discover causal relationships among variables from observational data. Learning the structures of causal networks provides an approach for causal relationship discovery. In many applications, we may be interested in finding the causes and effects of a given variable rather than a whole causal network over all variables. We reviewed several local structure learning approaches that find the local causal network around the variable of interest from observed data. These local structure learning algorithms can also be used to design an observational study before collecting data.

While causal effect evaluation focuses on identifying the effects of causes, the problem of causes of effects is of much interest in many legal and scientific studies. Dawid et al. (2014, 2015, 2016)

discussed the relationships and the differences between these two concepts and provided a framing for assessing causes of effects.

SUMMARY POINTS

1. The Yule-Simpson paradox means that the association between two variables can be changed by omitting a confounder. The criteria for confounders can be used for designs of observational studies and for data analysis.
2. Approaches to adjusting for observed confounders include stratification, matching, inverse probability weighting, regression, and hybrids of these approaches. For the case with unobserved confounders, instrumental variables are widely used, and the negative control approach is a promising mitigation method in the case where the instrumental variable approach fails.
3. The surrogate paradox is when a treatment has a positive causal effect on a surrogate, and the surrogate has a positive causal effect on an endpoint, but the treatment may have a negative causal effect on the endpoint. To avoid the surrogate paradox, criteria for surrogates can be used.
4. To find the causes and effects of the target variable, we need to learn only the local structure around the target variable rather than a global causal network. The MB-by-MB algorithm is reviewed.

FUTURE ISSUES

1. New criteria for surrogates need to be explored further to avoid the surrogate paradox and to quantitatively predict the causal effect of treatment on endpoints.
2. For causal network learning, approaches for dealing with latent variables need to be further developed.
3. In modern data science, causal inference will involve the integration of multisource data and summary data. The approaches of meta-analysis and transfer learning from different populations offer the potential to develop novel methods for causal inference.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to thank Nancy Reid and the Editorial Committee for encouraging us to write this review article, and we would also like to thank Anne Abramson and the reviewer for very helpful comments that have significantly improved the article. This research was supported by 973 Program of China (2015CB856000) and NSFC (11331011, 11771028, 91630314).

LITERATURE CITED

Abadie A, Imbens GW. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74:235–67

Aliferis C, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD. 2010. Local causal and Markov blanket induction for causal discovery and feature selection. *J. Mach. Learn. Res.* 11:235–84

Alonso A, Molenbergh G. 2008. Surrogate end points: hopes and perils. *Expert Rev. Pharmacoecon. Outcomes Res.* 8:255–59

Andersson SA, Madigan D, Perlman MD. 1997. A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Stat.* 25:505–41

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91:444–55

Bai X, Glymour C, Padman R, Ramsey J, Spirtes P, Wimberly F. 2004. *PCX: Markov blanket classification for large data sets with few cases*. Tech. Rep. CMU-CALD-04-102, Sch. Comput. Sci., Carnegie Mellon Univ.

Bai X, Padman R, Ramsey J, Spirtes P. 2008. Tabu search-enhanced graphical models for classification in high dimensions. *INFORMS J. Comput.* 20:423–37

Baker S. 2006. Surrogate endpoints: wishful thinking or reality? *J. Natl. Cancer Inst.* 98:502–3

Balke A, Pearl J. 1997. Bounds on treatment effects from studies with imperfect compliance. *J. Am. Stat. Assoc.* 92:1171–76

Birch MW. 1963. Maximum likelihood in three-way contingency tables. *J. R. Stat. Soc. B* 25:220–23

Bishop YMM, Fienberg SE, Holland PW. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press

Bound J, Jaeger DA, Baker RM. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Am. Stat. Assoc.* 90:443–50

Bowden J, Davey Smith G, Burgess S. 2015. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44:512–25

Breslow NE, Day NE. 1980. The analysis of case-control studies. *Annu. Rev. Public Health* 3:29–54

Brockwell SE, Gordon IR. 2001. A comparison of statistical methods for meta-analysis. *Stat. Med.* 20:825–40

Burzykowski T, Molenberghs G, Buyse M. 2005. *The Evaluation of Surrogate Endpoints*. New York: Springer

Buyse M, Molenberghs G. 1998. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 54:1014–29

CAST (The Cardiac Arrhythmia Suppression Trial Investigators). 1989. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N. Engl. J. Med.* 321:406–12

Chen H, Geng Z, Jia J. 2007. Criteria for surrogate end points. *J. R. Stat. Soc. B* 69:919–32

Clarke PS, Windmeijer F. 2012. Instrumental variable estimators for binary outcomes. *J. Am. Stat. Assoc.* 107:1638–52

Cochran WG, Rubin DB. 1973. Controlling bias in observational studies: a review. *Sankhya Ser. A* 35:417–46

Cooper GF, Yoo C. 1999. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, ed. KB Laskey, H Prade, pp. 116–25. San Francisco: Morgan Kaufmann

Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. 1959. Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* 22:173–203

Cox DR, Wermuth N. 2003. A general condition for avoiding effect reversal after marginalization. *J. R. Stat. Soc. B* 65:937–41

Crump RK, Hotz VJ, Imbens GW, Mitnik OA. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96:187–99

Darroch JN, Lauritzen SL, Speed TP. 1980. Markov fields and log-linear interaction models for contingency tables. *Ann. Stat.* 8:522–39

Davey Smith G. 2008. Assessing intrauterine influences on offspring health outcomes: Can epidemiological studies yield robust findings? *Basic Clin. Pharmacol.* 102:245–56

Dawid AP, Faigman DL, Fienberg SE. 2014. Fitting science into legal contexts: assessing effects of causes or causes of effects? *Sociol. Method. Res.* 43:359–90

Dawid AP, Faigman DL, Fienberg SE. 2015. On the causes of effects: Response to Pearl. *Sociol. Method. Res.* 44:165–74

Dawid AP, Musio M, Fienberg SE. 2016. From statistical evidence to evidence of causality. *Bayesian Anal.* 11:725–52

DerSimonian R, Laird N. 1986. Meta-analysis in clinical trials. *Control. Clin. Trials* 7:177–88
Drton M, Maathuis MH. 2017. Structure learning in graphical modeling. *Annu. Rev. Stat. Appl.* 4:365–93
Fienberg SE. 1980. *The Analysis of Cross-Classified Categorical Data*. Cambridge, MA: MIT Press
Fienberg SE. 2011. The analysis of contingency tables: from chi-squared tests and log-linear models to models of mixed membership. *Stat. Biopharm. Res.* 3:173–84
Flanders WD, Klein M, Darrow LA, Strickland MJ, Sarnat SE, et al. 2011. A method for detection of residual confounding in time-series and other observational studies. *Epidemiology* 22:59–67
Flanders WD, Strickland MJ, Klein M. 2017. A new method for partial correction of residual confounding in time-series and other observational studies. *Am. J. Epidemiol.* 185:941–49
Fleming TR, Demets DL. 1996. Surrogate end points in clinical trials: Are we being misled? *Ann. Intern. Med.* 125:605–13
Fogarty CB, Mikkelsen ME, Gaieski DF, Small DS. 2016. Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *J. Am. Stat. Assoc.* 111:447–58
Frangakis CE, Rubin DB. 1999. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* 86:365–79
Frangakis CE, Rubin DB. 2002. Principal stratification in causal inference. *Biometrics* 58:21–29
Freedman LS, Graubard BI, Schatzkin A. 1992. Statistical validation of intermediate endpoints for chronic diseases. *Stat. Med.* 11:167–78
Gagnon-Bartsch JA, Speed TP. 2012. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* 13:539–52
Gail MH. 1986. Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In *Modern Statistical Methods in Chronic Disease Epidemiology*, ed. SH Moolgavkar, RL Prentice, pp. 3–18. New York: Wiley
Geng Z. 1992. Collapsibility of relative risk in contingency tables with a response variable. *J. R. Stat. Soc. B* 54:585–93
Geng Z. 2015. Surrogates for qualitative evaluation of treatment effects. In *Advanced Medical Statistics*, ed. Y Lu, J Fang, L Tian, H Jin, pp. 781–802. London: World Sci.
Geng Z, Guo JH, Fung WK. 2002. Criteria for confounders in epidemiological studies. *J. R. Stat. Soc. B* 64:3–15
Geng Z, Guo JH, Tai SL, Fung W. 2001. Confounding, homogeneity and collapsibility for causal effects in epidemiologic studies. *Stat. Sinica* 11:63–75
Geng Z, Li G. 2002. Conditions for non-confounding and collapsibility without knowledge of completely constructed causal diagrams. *Scand. J. Stat.* 29:169–81
Gilbert PB, Hudgens MG. 2008. Evaluating candidate principal surrogate endpoints. *Biometrics* 64:1146–54
Goldberger AS. 1972. Structural equation methods in the social sciences. *Econometrica* 40:979–1001
Goodman LA. 1970. The multivariate analysis of qualitative data: interactions among multiple classifications. *J. Am. Stat. Assoc.* 65:226–56
Greenland S, Pearl J. 2011. Adjustments and their consequences—collapsibility analysis using graphical models. *Int. Stat. Rev.* 79:401–26
Greenland S, Pearl J, Robins JM. 1999a. Causal diagrams for epidemiologic research. *Epidemiology* 10:37–48
Greenland S, Robins JM. 1986. Identifiability, exchangeability, and epidemiological confounding. *Int. J. Epidemiol.* 15:413–19
Greenland S, Robins JM, Pearl J. 1999b. Confounding and collapsibility in causal inference. *Stat. Sci.* 14:29–46
Guyon I, Aliferis C, Cooper GF, Elisseeff A, Pellet JP, et al. 2008. Design and analysis of the causation and prediction challenge. In *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*, ed. I Guyon, C Aliferis, G Cooper, A Elisseeff, JP Pellet, et al., pp. 1–33. **http://proceedings.mlr.press/v3/guyon08a/guyon08a.pdf**
Haberman SJ. 1974. *The Analysis of Frequency Data*. Chicago: Univ. Chicago Press
Hahn J. 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66:315–31

Hauser A, Bühlmann P. 2012. Two optimal strategies for active learning of causal models from interventional data. *Int. J. Approx. Reason.* 55:926–39

He YB, Geng Z. 2008. Active learning of causal networks with intervention experiments and optimal designs. *J. Mach. Learn. Res.* 9:2523–47

Heckman J, Ichimura H, Smith J, Todd P. 1998. Characterizing selection bias using experimental data. *Econometrica* 66:1017–98

Heinze-Deml C, Maathuis MH, Meinshausen N. 2018. Causal structure learning. *Annu. Rev. Stat. Appl.* 5:371–91

Hernán MA, Robins JM. 2006. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 17:360–72

Hernán MA, Robins JM. 2018. *Causal Inference*. Boca Raton, FL: Chapman & Hall

Hill AB. 1965. The environment and disease: association or causation? *Proc. R. Soc. Med.* 58:295–300

Hill J, Su YS. 2013. Assessing lack of common support in causal inference using Bayesian nonparametrics: implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Ann. Appl. Stat.* 7:1386–420

Horvitz DG, Thompson DJ. 1952. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47:663–85

Imbens GW, Angrist JD. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62:467–75

Imbens GW, Rubin DB. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, UK: Cambridge Univ. Press

Inoue A, Solon G. 2010. Two-sample instrumental variables estimators. *Rev. Econ. Stat.* 92:557–61

Joffe M. 2013. Discussion on "Surrogate measures and consistent surrogates." *Biometrics* 69:572–75

Joffe M, Greene T. 2009. Related causal frameworks for surrogate outcomes. *Biometrics* 65:530–38

Ju C, Geng Z. 2010. Criteria for surrogate end points based on causal distributions. *J. R. Stat. Soc. B* 72:129–42

Kang H, Zhang A, Cai TT, Small DS. 2016. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *J. Am. Stat. Assoc.* 111:132–44

Kang JD, Schafer JL. 2007. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat. Sci* 22:523–39

King G, Nielsen R. 2016. *Why propensity scores should not be used for matching*. Work. Pap., Harvard Univ. **https://gking.harvard.edu/publications/why-Propensity-Scores-Should-Not-Be-Used-Formatching**

Kolesár M, Chetty R, Friedman J, Glaeser E, Imbens GW. 2015. Identification and inference with many invalid instruments. *J. Bus. Econ. Stat.* 33:474–84

Lauritzen S. 2004. Discussion on causality. *Scand. J. Stat.* 31:189–93

Lee BK, Lessler J, Stuart EA. 2010. Improving propensity score weighting using machine learning. *Stat. Med.* 29:337–46

Li Y, Taylor J, Elliott MR. 2010. A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics* 66:523–31

Lin W, Feng R, Li H. 2015. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *J. Am. Stat. Assoc.* 110:270–88

Lipsitch M, Tchetgen Tchetgen E, Cohen T. 2010. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 21:383–88

Luo P, Cai Z, Geng Z. 2018. Criteria for multiple surrogates. *Stat. Sinica*. In press

Ma Z, Xie X, Geng Z. 2006. Collapsibility of distribution dependence. *J. R. Stat. Soc. B* 68:127–33

Manns B, Owen WF, Winkelmayer WC, Devereaux PJ, Tonelli M. 2006. Surrogate markers in clinical studies: problems solved or created? *Am. J. Kidney Dis.* 48:159–66

Manski CF. 1990. Nonparametric bounds on treatment effects. *Am. Econ. Rev.* 80:319–23

Manski CF, Pepper JV. 2000. Monotone instrumental variables with an application to the returns to schooling. *Econometrica* 68:997–1010

Mantel N. 1989. Confounding in epidemiologic studies. *Biometrics* 45:1317–18

Mantel N, Haenszel W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* 22:719–48

Miao W, Geng Z, Tchetgen Tchetgen E. 2018. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* 105:987–93
Miao W, Tchetgen Tchetgen E. 2017. Invited commentary: bias attenuation and identification of causal effects with multiple negative controls. *Am. J. Epidemiol.* 185:950–53
Miettinen OS, Cook EF. 1981. Confounding: essence and detection. *Am. J. Epidemiol.* 114:593–603
Moore T. 1995. *Deadly Medicine: Why Tens of Thousands of Patients Died in America's Worst Drug Disaster*. New York: Simon & Schuster
Murphy KP. 2001. *Active learning of causal Bayes net structure*. Work. Pap., Dep. Comput. Sci., Univ. Calif., Berkeley
Pearl J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann
Pearl J. 1995. Causal diagrams for empirical research. *Biometrika* 82:669–88
Pearl J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge Univ. Press. 2nd ed.
Peña JM, Nilsson R, Björkegren J, Tegnér J. 2007. Towards scalable and data efficient learning of Markov boundaries. *Int. J. Approx. Reason.* 45:211–32
Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. 2012. Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* 21:31–54
Pocock SJ, Assmann SE, Enos LE, Kasten LE. 2002. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat. Med.* 21:2917–30
Prentice RL. 1989. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat. Med.* 8:431
Ramsey J. 2006. *A PC-style Markov blanket search for high dimensional datasets*. Tech. Rep. CMU-PHIL-177, Dep. Philosophy, Carnegie Mellon Univ.
Robins JM. 1994. Correcting for non-compliance in randomized trials using structural nested mean models. *Commun. Stat. Theor. Methods* 23:2379–412
Robins JM, Mark SD, Newey WK. 1992. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 48:479–95
Robins JM, Rotnitzky A, Zhao LP. 1994. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.* 89:846–66
Robinson LD, Jewell NP. 1991. Some surprising results about covariate adjustment in logistic regression models. *Int. Stat. Rev.* 59:227–40
Rosenbaum PR. 1989. The role of known effects in observational studies. *Biometrics* 45:557–69
Rosenbaum PR. 2002. *Observational Studies*. New York: Springer. 2nd ed.
Rosenbaum PR, Rubin DB. 1983a. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Stat. Soc. B* 45:212–18
Rosenbaum PR, Rubin DB. 1983b. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55
Rubin DB. 1973. Matching to remove bias in observational studies. *Biometrics* 29:159–83
Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6:34–58
Rubin DB. 1980. Randomization analysis of experimental data: the Fisher randomization test: comment. *J. Am. Stat. Assoc.* 75:591–93
Rubin DB. 2008. For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* 2:808–40
Scharfstein DO, Rotnitzky A, Robins JM. 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Am. Stat. Assoc.* 94:1096–120
Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. 2014. Interpreting observational studies: why empirical calibration is needed to correct *p*-values. *Stat. Med.* 33:209–18
Simpson EH. 1951. The interpretation of interaction in contingency tables. *J. R. Stat. Soc. B* 13:238–41
Small DS. 2007. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J. Am. Stat. Assoc.* 102:1049–58
Stock JH, Wright JH, Yogo M. 2002. A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econ. Stat* 20:518–29
Stuart EA. 2010. Matching methods for causal inference: a review and a look forward. *Stat. Sci.* 25:1–21
Tian J, Pearl J. 2001. Causal discovery from changes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, ed. JS Breese, D Koller, pp. 512–21. San Francisco: Morgan Kaufmann

Tong S, Koller D. 2001. Active learning for structure in Bayesian networks. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 863–69. Palo Alto, CA: AAAI

Trichopoulos D, Zavitsanos X, Katsouyanni K, Tzonou A, Dalla-Vorgia P. 1983. Psychological stress and fatal heart attack: the Athens (1981) earthquake natural experiment. *Lancet* 321:441–44

VanderWeele TJ. 2013. Surrogate measures and consistent surrogates. *Biometrics* 69:561–81

VanderWeele TJ, Shpitser I. 2011. A new criterion for confounder selection. *Biometrics* 67:1406–13

Verma T, Pearl J. 1990. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, ed. PP Bonissone, M Henrion, LN Kanal, JF Lemmer, pp. 220–27. San Francisco: Morgan Kaufmann

Wang CZ, Zhou Y, Zhao Q, Geng Z. 2014. Discovering and orienting the edges connected to a target variable in a DAG via a sequential local learning approach. *Comput. Stat. Data Anal.* 77:252–66

Wang J, Zhao Q, Hastie T, Owen AB. 2017. Confounder adjustment in multiple hypothesis testing. *Ann. Stat.* 45:1863–94

Wang L, Robins JM, Richardson TS. 2017. On falsification of the binary instrumental variable model. *Biometrika* 104:229–36

Wang X, Geng Z, Chen H, Xie X. 2009. Detecting multiple confounders. *J. Stat. Plan. Inference* 139:1073–81

Wang X, Geng Z, Zhao Q, Qiao Q. 2007. Comparison between estimates of the hypothetical proportion with and without standardization for a non-confounder. *Stat. Sinica* 17:91–93

Weiss NS. 2002. Can the specificity of an association be rehabilitated as a basis for supporting a causal hypothesis? *Epidemiology* 13:6–8

Whittemore AS. 1978. Collapsibility of multidimensional contingency tables. *J. R. Stat. Soc. B* 40:328–40

Wickramaratne PJ, Holford TR. 1987. Confounding in epidemiologic studies: the adequacy of the control group as a measure of confounding. *Biometrics* 43:751–65

Wickramaratne PJ, Holford TR. 1989. Confounding in epidemiologic studies. Response. *Biometrics* 45:1319–22

Wright PG. 1928. *Tariff on Animal and Vegetable Oils*. New York: Macmillan

Wu Z, He P, Geng Z. 2011. Sufficient conditions for concluding surrogacy based on observed data. *Stat. Med.* 30:2422–34

Xie X, Geng Z. 2008. A recursive method for structural learning of directed acyclic graphs. *J. Mach. Learn. Res.* 9:459–83

Xie X, Geng Z, Zhao Q. 2006. Decomposition of structural learning about directed acyclic graphs. *Artif. Intell.* 170:422–39

Xie X, Ma Z, Geng Z. 2008. Some association measures and their collapsibility. *Stat. Sinica* 18:1165–83

Yerushalmy J, Palmer CE. 1959. On the methodology of investigations of etiologic factors in chronic diseases. *J. Chron. Dis.* 10:27–40

Yule GU. 1903. Notes on the theory of association of attributes in statistics. *Biometrika* 2:121–34