ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Sibling Comparison Studies

Arvid Sjölander,[1] Thomas Frisell,[2] and Sara Öberg[1]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden 171 77; email: arvid.sjolander@ki.se

[2]Department of Medicine, Karolinska Institute, Stockholm, Sweden 171 77

Annu. Rev. Stat. Appl. 2022. 9:71–94

First published as a Review in Advance on October 21, 2021

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

https://doi.org/10.1146/annurev-statistics-040120-024521

Keywords

between-within models, bias, causal inference, conditional effects, confounding, fixed effects models, marginal effects, random effects models, siblings, twins

Abstract

Unmeasured confounding is one of the main sources of bias in observational studies. A popular way to reduce confounding bias is to use sibling comparisons, which implicitly adjust for several factors in the early environment or upbringing without requiring them to be measured or known. In this article we provide a broad exposition of the statistical analysis methods for sibling comparison studies. We further discuss a number of methodological challenges that arise in sibling comparison studies.

1. INTRODUCTION

A common research objective is to estimate the causal effect of a particular exposure on a particular outcome. A major obstacle is that, unless the exposure is controlled and randomized by the researcher, there are often common causes—confounders—of the exposure and the outcome. In the presence of confounding, the exposure and the outcome are statistically associated even in the absence of a causal exposure effect.

To reduce confounding bias, researchers typically attempt to measure potential confounders and adjust for these in the statistical analysis. However, this strategy is often hampered by the facts that the number of potential confounders may be very large and that some important confounders may be hard to measure or even unknown by the researcher. A popular way to adjust for unmeasured confounders is to use designs that compare differentially exposed siblings. Since siblings are naturally matched on many potential confounders, including several factors in the early environment or upbringing, such sibling comparisons implicitly adjust for these shared confounders without requiring them to be measured or known. In this way, sibling comparison designs are distinct from more standard matched designs, where all matching variables are completely measured and determined by the investigator (Rosenbaum 2020). Rather, sibling comparison designs may be viewed as belonging to the more general class of quasi-experimental designs, which aim to adjust for unmeasured confounders as well (Rosenbaum 2015). A prominent special case is the co-twin control study, which, if restricted to genetically identical (i.e., monozygotic) twins, eliminates confounding by all heritable genetic factors.

The use of sibling comparisons dates back to at least the end of the nineteenth century, when Sullivan (1899) used siblings to estimate the effect of maternal alcoholism on offspring mortality. A few decades later, Gorseline (1932) used siblings to estimate the effect of education on salary. In the beginning of the twentieth century, the Medico-Biological Institute in the Soviet Union systematically collected data on twin pairs for various studies, of which some used a co-twin control design (Levit 1935). Other early studies with co-twin control designs were described by Gesell (1942). Notably though, in these studies the exposure (motor training) was randomized within twin pairs, which, in large samples, eliminates any confounding bias and thus makes the co-twin control design somewhat superfluous. From the mid-1900s, sibling/twin comparisons have been used extensively to study the effects of various exposures, such as smoking (Floderus et al. 1988, Piirtola et al. 2018), alcohol consumption (Lown et al. 2008, Dai et al. 2015), overweight and obesity (Jonsson et al. 2003, Boone-Heinonen et al. 2020), poor fetal growth (Lawlor et al. 2006, Class et al. 2014), stressful or traumatic life events (Eisen et al. 1991, Kendler et al. 1999), low cognitive ability (Murray 2002, Kolk & Barclay 2019), advanced parental age (Lawlor et al. 2011, D'Onofrio et al. 2014) and neurodevelopmental disorders (Lundström et al. 2014, Daley et al. 2019). In the Nordic countries, co-twin control studies are facilitated by the existence of nationwide twin registries (Skytthe et al. 2011, Nilsen et al. 2013, Zagai et al. 2019).

The aim of this article is to provide a broad exposition of the statistical analysis methods for sibling comparison studies. These have been developed with contributions from several fields, such as epidemiology, biostatistics, econometrics, social science, and causal inference. For ease of exposition, we mainly focus on outcomes that are measured at a single point in time, without truncation and censoring. However, most of the methods and models for point outcomes that we review have close analogies for time-to-event outcomes; we will indicate this as we proceed. We emphasize that sibling data are a special type of clustered data. Hence, several of the papers and books that we cite were primarily concerned with other types of clustered data, such as complex survey data (e.g., Cai & Brumback 2015), studies with repeated measures (e.g., Allison 2009), or just clustered data in general. A central feature of the methods and estimators that we review is their

asymptotic (i.e., large sample) bias with respect to a particular target parameter. For simplicity, we use the term bias throughout as shorthand for asymptotic bias.

Our focus is on estimating the causal effect of an exposure of interest and using siblings to adjust for unmeasured confounding. We note that siblings are often used for the somewhat different purpose of estimating the heritability of traits. By contrasting the correlation in a trait for individuals with different degrees of genetic and environmental relatedness (e.g., twins reared together, twins reared apart, full siblings, half siblings, cousins), it is possible to estimate the genetic contribution to the trait (Falconer & Mackay 1996). This topic is beyond the scope of our review, though.

The article is organized as follows. In Section 2, we introduce basic notation and definitions. In Section 3, we review the analysis methods for sibling comparison studies. We start this section with an account of model-free methods and then proceed to model-based methods. We end the section with a brief review of existing software that implements these methods. In Section 4, we discuss a number of methodological challenges that arise in sibling comparison studies. Finally, in Section 5, we illustrate the methods and concepts with an application to fetal growth restriction and attention-deficit/hyperactivity disorder (ADHD).

2. NOTATION AND DEFINITIONS

Let X_{ij} and Y_{ij} denote the exposure and outcome of interest, respectively, for sibling j within family i. Let C_i denote the set of shared confounders, i.e., those confounders that have the same value for all siblings in the same family. Let C_{ij} denote the set of nonshared confounders for sibling j within family i, i.e., those confounders that may vary across siblings within the same family. To distinguish between measured and unmeasured nonshared confounders, we use C_{ij}^m for the former. In practice, C_{ij}^m would often be a vector of variables. However, to simplify notation we assume that C_{ij}^m is a scalar, with obvious generalization to vectors. We assume that data are measured on siblings from n independent families, with n_i siblings in family i. Finally, for any scalar variable V_{ij}, we define the vector $\mathbf{V}_i = (V_{i1}, \ldots, V_{in_i})$ and the mean $\overline{V}_{i\cdot} = \sum_j V_{ij}/n_i$, for each family i.

The causal diagram (Pearl 1995, 2009) in **Figure 1** illustrates the situation. Although useful for pedagogical purposes, this causal diagram makes several simplifying assumptions; we problematize some of these assumptions in Section 4. The aim is to estimate the causal effect of the exposure on the outcome, represented by the arrow from X_{ij} to Y_{ij}. Let $Y_{ij}(x)$ be the potential outcome (Rubin 1974, Little & Rubin 2000) for sibling j in family i that we would have observed, had that subject been exposed to $X_{ij} = x$. Causal effects are defined as contrasts (e.g., mean differences or risk ratios) between potential outcomes (Pearl 2009, Hernán & Robins 2020). The sibling comparison methods that we review in the subsequent sections implicitly adjust for all shared confounders, regardless of whether these are measured or not. However, these methods do not adjust for unmeasured nonshared confounders. Hence, to make a causal interpretation of estimates obtained from sibling comparisons, one has to assume that there are no unmeasured nonshared confounders.

As discussed in Section 4, sibling comparison studies mainly use information from families where there is variation in the exposure, and they largely ignore families with no variation in the exposure; we refer to siblings from these families as exposure discordant and exposure concordant, respectively. When the exposure is binary and there are two siblings in each family, the exposure-discordant pairs are those where one sibling is exposed ($X = 1$) and one sibling is unexposed ($X = 0$), and the exposure-concordant pairs are those where both siblings are either exposed or unexposed.

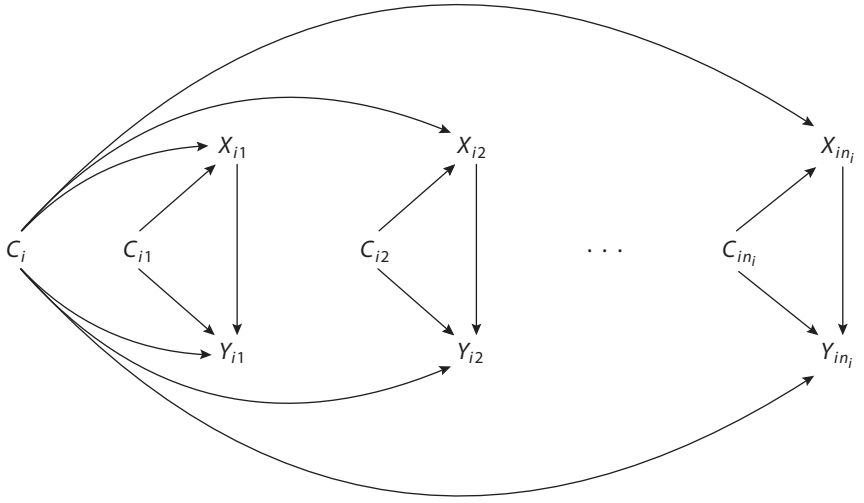

Figure 1

Causal diagram illustrating a family with n_i siblings in a sibling comparison study. The variables X_{ij}, Y_{ij}, and C_{ij} represent the exposure, outcome, and nonshared confounders, respectively, for sibling j within family i, and the variable C_i represents the set of shared confounders.

3. STATISTICAL ANALYSIS METHODS

3.1. Model-Free Analysis

As a starting point, suppose that the exposure is binary (0/1) and that all families have exactly two siblings (i.e., a sib-pair), which is always the case in co-twin control studies. In this special case we can adjust for the shared confounders C_i without making any parametric model assumptions. To this end, we restrict attention to the exposure-discordant pairs and compare the distribution of the outcome between the exposed and the unexposed among these pairs. Since C_i is constant within each sib-pair, the restriction to exposure-discordant pairs ensures that, for each value of C_i, there are exactly the same number of exposed and unexposed subjects. For instance, if C_i has a unique value for each sib-pair, then there will be exactly one exposed and one unexposed for each value of C_i in the restricted sample of exposure-discordant pairs. Or, to put it the other way around, in this restricted sample, the exposed and the unexposed have identical distributions of C_i. Hence, if we observe that the outcome distribution differs between exposed and unexposed in the exposure-discordant pairs, then we cannot attribute this to systematic differences in (i.e., confounding by) C_i.

Comparing the outcome distribution between exposed and unexposed does not require any parametric model assumptions and can be carried out with standard analytic methods. For instance, to test for an exposure effect, we may use a standard test for paired data, e.g., McNemar's test (for binary outcomes), the Wilcoxon signed-rank test (for continuous outcomes), or a paired log-rank test (for time-to-event outcomes; Jung 1999). To estimate the exposure effect, we may compute any desired contrast between the exposed and the unexposed, such as the risk difference, the risk ratio or the odds ratio (for binary outcomes), or the mean difference (for continuous outcomes). For time-to-event outcomes, we may compute and compare the Kaplan–Meier curves (Kaplan & Meier 1958, Klein & Moeschberger 2003) for the exposed and the unexposed. To assess the sampling variability in the estimate, we would typically like to provide a confidence interval for the true effect. A simple and general method is to use a Wald confidence interval on the form of

estimate ± standard error, where the standard error may be computed with the sandwich formula (Stefanski & Boos 2002) to account for the paired data structure; Sjölander et al. (2012b, appendix A) provide a worked example for binary outcomes.

However simple, this analysis has the usual limitations of model-free analyses, such as potentially low statistical power, difficulty in accommodating nonbinary exposures, and difficulty in adjusting for measured confounders. In addition, the model-free analysis does not generalize easily to families with more than two siblings. A more general and flexible analysis is based on regression models. There is a plethora of model-based analysis methods for sibling comparison studies. To provide a structured exposition and to illustrate how different methods relate to each other, we organize the methods according to what parameter they aim to estimate and what parametric assumptions they make. We distinguish between parameters that, in the absence of unmeasured nonshared confounders, can be interpreted as conditional causal effects and marginal causal effects. We further distinguish between models that parameterize the relation between the confounders and the outcome, and models that parameterize the relation between the confounders and the exposure; we refer to these different models as outcome regressions and exposure regressions, respectively. Estimation of conditional effects with outcome regression is, by far, the most common modeling strategy. However, marginal effects and exposure regression have gained popularity in recent years, particularly with the influence from the causal inference field.

3.2. Model-Based Estimation of Conditional Causal Effects

In this section, we review the methods for model-based estimation of conditional causal effects.

3.2.1. Outcome regression. To analyze sibling data, it is common to use fixed effects models of the form

$$g\{\mathrm{E}(Y_{ij}|C_i, C_{ij}^m, X_{ij})\} = \alpha_i + \gamma C_{ij}^m + \beta X_{ij}, \qquad 1.$$

where $\mathrm{E}(Y_{ij}|\cdot)$ is the conditional mean of the outcome and g is an appropriate link function, typically the identity link, the log link, or the logit link. The term fixed here refers to the intercept α_i, which is considered a categorical parameter with one fixed level per family. This intercept is intended to absorb, and thereby adjust for, the shared confounders C_i. An analogous fixed effects model for time-to-event outcomes is the Cox proportional hazards model with a family-specific baseline hazard (Holt & Prentice 1974).

Model 1, as well as all subsequent models, assumes that there are no interactions between $(\alpha_i, X_{ij}, C_{ij}^m)$ and that all effects are linear, on the scale defined by g. Assuming no interactions between α_i and (X_{ij}, C_{ij}^m) is necessary, since such interactions pose identifiability problems for the estimators that we consider (Zetterqvist et al. 2016). The remaining assumptions are mainly to keep notation simple; in practice, we may allow for more complex relations by adding interactions between X_{ij} and C_{ij}^m and/or nonlinear effects of these, such as splines. However, we note that more elaborate models also make interpretation more difficult. We return to these important modeling issues in Sections 3.4 and 4.2.

In Model 1, the exposure coefficient β measures the conditional association of the outcome with one unit increase in the exposure:

$$\beta = g\{\mathrm{E}(Y_{ij}|C_i, C_{ij}^m, X_{ij} = x + 1)\} - g\{\mathrm{E}(Y_{ij}|C_i, C_{ij}^m, X_{ij} = x)\}. \qquad 2.$$

When there are no unmeasured nonshared confounders, β can be interpreted as the conditional causal effect

$$g[\mathrm{E}\{Y_{ij}(x+1)|(C_i, C_{ij})\}] - g[\mathrm{E}\{Y_{ij}(x)|(C_i, C_{ij})\}], \quad 3.$$

i.e., the conditional effect of increasing the exposure with one unit, given (C_i, C_{ij}).

A naive approach to estimate β would be to assume a parametric distribution for Y_{ij} and use standard maximum likelihood (ML) to estimate β jointly with γ and α_i. Since α_i has one level, and thus one parameter, per family, this approach can be computationally demanding. More seriously, apart from some special cases, it gives an inconsistent estimate of β (Allison 2009). The reason for this is that standard ML estimation fails when the number of parameters in the model increases with the sample size (e.g., the number of families); this is often referred to as the incidental variable problem (Lancaster 2000). To bypass this problem, β is usually estimated with conditional ML, which eliminates α_i by conditioning on the sufficient statistic $\sum_j Y_{ij}$ for each family i, thereby producing a consistent estimator of β (Andersen 1970, Neuhaus & McCulloch 2006). When Y_{ij} is binary and g is the logit link, this is referred to as conditional logistic regression (Breslow & Day 1980, chapter 7). Goetgeluk & Vansteelandt (2008) proposed an alternative semiparametric estimator, which requires g to be the identity link or log link but does not require an assumed parametric distribution for Y_{ij}. The Cox proportional hazards model with a family-specific baseline hazard can be fitted with partial likelihood methods; this is referred to as stratified Cox regression (Klein & Moeschberger 2003, chapter 9.3).

An alternative approach is to use a random effects model. In this approach, α_i is assumed to have a parametric (e.g., normal) distribution, and β is estimated by integrating out α_i from the likelihood. However, this approach is not suitable for sibling comparison studies. This is because the standard formulation of the random effects model additionally assumes that α_i is statistically independent of the covariates in the model, e.g., X_{ij} and C_{ij}; this assumption is implicit in standard software such as the `glmer` function in R; the `GLIMMIX` procedure in SAS; and the `xtreg`, `xtpoisson`, and `xtlogit` commands in Stata. Since α_i is supposed to absorb the shared confounders C_i, and since a confounder by definition has to be associated with the exposure, the model thus assumes that there are no shared confounders. When this assumption does not hold, which is typically the case in sibling comparison studies, the model fails to adjust for the shared confounders and thus gives biased estimates (Allison 2009).

Yet another alternative, which does not suffer from the problem of standard random effects models, is to use a so-called between-within (BW) model (Mundlak 1978, Neuhaus & McCulloch 2006), also referred to as a hybrid model (Allison 2009) or a poor man's approximation (to the fixed effects model) (Neuhaus & McCulloch 2006, Brumback et al. 2010). To motivate this model, note that we may allow for α_i to depend on $\mathbf{X}_i$ and $\mathbf{C}_i^m$ by assuming that

$$\alpha_i = \alpha_i^\dagger + \gamma_B \overline{C}_{i\cdot}^m + \beta_B \overline{X}_{i\cdot}, \quad 4.$$

where $\alpha_i^\dagger$ has a $\mathrm{N}(\mu, \sigma^2)$ distribution and is statistically independent of $(\mathbf{X}_i, \mathbf{C}_i^m)$. Combining Model 1 with the additional assumption in Equation 4 gives the equivalent model formulation

$$g\{\mathrm{E}(Y_{ij}|C_i, C_{ij}^m, X_{ij})\} = \alpha_i^\dagger + \gamma_B \overline{C}_{i\cdot}^m + \beta_B \overline{X}_{i\cdot} + \gamma C_{ij}^m + \beta X_{ij}, \quad 5.$$

Here, the parameters (γ_B, β_B) and (γ, β) are referred to as between-effects and within-effects, respectively. Another common formulation replaces the terms γC_{ij}^m and βX_{ij} in the model with

$\gamma(C_{ij}^m - \overline{C}_{i\cdot}^m)$ and $\beta(X_{ij} - \overline{X}_{i\cdot})$, respectively. This formulation reparameterizes the between-effects into $\gamma_B - \gamma$ and $\beta_B - \beta$ but leaves the within-effects unchanged. An analogous BW model was proposed for time-to-event outcomes by Sjölander et al. (2013).

From its derivation through Equation 4, it is clear that the BW model allows for nonshared confounders by allowing for α_i to depend on $\mathbf{X}_i$ (and $\mathbf{C}_i^m$). However, the formulation of the model in Equation 5 is still in the standard random effects model format, with an intercept that is independent of the model covariates, which means that the model can be fitted with standard software for random effects models.

The BW model makes stronger parametric assumptions than the fixed effects model, in that it assumes a parametric distribution for α_i. Hence, one would hope that the BW model also gives more efficient estimates. Unfortunately, though, this is typically not the case. It can be shown that the fixed effects model and the BW model give identical estimates of β when g is the identity link (Seaman et al. 2014). For other link functions, the estimates are generally not identical but often very similar (Neuhaus & McCulloch 2006). This indicates that the BW model does not, in general, provide any efficiency gain over the fixed effects model. However, the similarity of the estimates also indicates that the BW model is fairly robust against its additional assumptions and that it may often give consistent estimates even if these assumptions are wrong. This is not guaranteed, though; Goetgeluk & Vansteelandt (2008) and Brumback et al. (2010) gave numerical counterexamples, showing that incorrectly specified BW models with log links and logit links, respectively, may occasionally give biased estimates. Sjölander et al. (2013) showed by simulation that the BW model for time-to-event outcomes behaves somewhat differently than the BW model for point outcomes, in that it often gives more efficient estimates than the corresponding fixed effects (i.e., stratified Cox) model. However, these authors also gave numerical examples showing that the efficiency gain may come at the price of biased estimates, if the additional assumptions of the BW model are wrong.

A potential advantage of the BW model is that it provides a way of quantifying the degree of shared confounding. In the complete absence of shared confounding (i.e., when C_i is empty), we would expect that α_i is independent of $\mathbf{X}_i$, so that $\beta_B = 0$. Conversely, an estimate of β_B that differs significantly from 0 signals the presence of shared confounding. Intuitively, then, we can use β_B as a measure of the degree of shared confounding, where stronger deviations from 0 indicate a higher degree of confounding. However, an important disadvantage of the BW model is that it generally requires numeric approximation of complex likelihood integrals and is thus more computationally demanding than the fixed effects model (e.g., Sjölander 2021).

3.2.2. Exposure regression. Both the fixed effects model in Equation 1 and the BW model in Equation 5 are outcome regressions, in the sense that they model how the outcome depends on the shared confounders and the measured nonshared confounders. Specifically, both models can be partitioned into the target parameter β, as defined in Equation 2, and the outcome nuisance model,

$$g\{\mathrm{E}(Y_{ij}|C_i, C_{ij}^m, X_{ij} = 0)\} = \alpha_i + \gamma C_{ij}^m. \quad 6.$$

This nuisance model essentially parameterizes the arrows from (C_i, C_{ij}) to Y_{ij} in **Figure 1**. In some scenarios, though, the researcher may prefer to use a model that parameterizes the arrows from (C_i, C_{ij}) to X_{ij}.

Zetterqvist et al. (2016) showed how the target parameter β can be estimated with a regression model for the exposure, when the link function g in Equation 2 is either the identity link or the log

link. They replaced the outcome nuisance model with an exposure nuisance model on the form

$$g^*\{\mathrm{E}(X_{ij}|C_i, C_{ij}^m)\} = \alpha_i^* + \gamma^* C_{ij}^m, \quad 7.$$

where we have used superindex* to distinguish between link functions and parameters in the models in Equations 6 and 7. They derived an unbiased estimating equation for β that depends on the nuisance parameter γ^* and proposed to estimate β by solving this estimating equation, with γ^* replaced by a consistent estimate thereof. Since the exposure nuisance model in Equation 7 is a standard fixed effects model, a consistent estimate of γ^* may be obtained either with conditional ML, or with the semiparametric method of Goetgeluk & Vansteelandt (2008) if g^* is the identity or log link.

Zetterqvist et al. (2019) showed how an analogous estimator of β can be obtained in the special case when both X_{ij} and Y_{ij} are binary and both g and g^* are logit links. This estimator uses the fact that, due to the symmetry of the odds ratio, β can, in this special case, also be formulated as

$$\beta = \mathrm{logit}\{\mathrm{E}(X_{ij}|C_i, C_{ij}^m, Y_{ij} = 1)\} - \mathrm{logit}\{\mathrm{E}(X_{ij}|C_i, C_{ij}^m, Y_{ij} = 0)\}. \quad 8.$$

Combining Equation 8 with the exposure nuisance model in Equation 7 gives the conditional logistic regression model

$$\mathrm{logit}\{\mathrm{E}(X_{ij}|C_i, C_{ij}^m, Y_{ij})\} = \alpha_i^* + \gamma^* C_{ij}^m + \beta Y_{ij}. \quad 9.$$

A consistent estimate of β can be obtained from the model in Equation 9 using conditional ML; Zetterqvist et al. (2019) referred to this as retrospective conditional logistic regression.

In many scenarios, the researcher may not have a clear preference for the model in either Equation 6 or 7. In such scenarios it could be desirable to have a doubly robust (DR) estimator of β, i.e., an estimator that uses both a nuisance model for the outcome and a nuisance model for the exposure, and is consistent if either model is correct, not necessarily both (Bang & Robins 2005). Zetterqvist et al. (2016, 2019) derived such DR estimators for β as well.

3.3. Model-Based Estimation of Marginal Causal Effects

The causal effect in Equation 3 is conditional, in the sense that it applies to groups defined by fixed values of the confounders (C_i, C_{ij}). However, in some situations it may be more desirable to estimate a population effect that applies marginally over the confounders. Before discussing how this can be done with regression models, we note that the model-free analysis described in Section 3.1 can be viewed as attempting to estimate one such marginal effect. When contrasting the mean outcome for exposed and unexposed among exposure-discordant pairs, we are in effect estimating

$$h\{\mathrm{E}(Y_{ij}|X_{ij} = 1, X_{i1} \neq X_{i2})\} - h\{\mathrm{E}(Y_{ij}|X_{ij} = 0, X_{i1} \neq X_{i2})\}, \quad 10.$$

where h is a link function that defines our contrast, e.g., identity, log, or logit. For instance, when Y_{ij} is binary and h is the logit link, the contrast in Equation 10 is the exposure-outcome log odds ratio among the exposure-discordant pairs. When there are no unmeasured nonshared confounders, it can be shown (Sjölander et al. 2012b) that the contrast in Equation 10 can be interpreted as the

causal effect

$$h[\mathrm{E}^d\{Y_{ij}(1)\}] - h[\mathrm{E}^d\{Y_{ij}(0)\}]. \quad 11.$$

In this expression, the expectations are taken with respect to a potential outcome distribution where C_i is distributed as among the exposure-discordant pairs: $\mathrm{E}^d\{Y_{ij}(x)\} = \mathrm{E}[\mathrm{E}\{Y_{ij}(x)|C_{ij}\}|X_{i1} \neq X_{i2}]$. Hence, we may interpret the effect in Equation 11 as the marginal (over C_i) causal effect among the exposure-discordant pairs. In the next section we show how we can use regression modeling to estimate the effect among all pairs.

3.3.1. Outcome regression. We define the marginal causal effect in the whole sample (e.g., both exposure-discordant and exposure-concordant pairs) as

$$h[\mathrm{E}\{Y_{ij}(x')\}] - h[\mathrm{E}\{Y_{ij}(x)\}]. \quad 12.$$

In this expression, the expectations are taken with respect to the distribution of C_i (and C_{ij}) in the whole sample, e.g., the mix of exposure-discordant and exposure-concordant pairs. The values x' and x are two exposure levels (e.g., 1 and 0) that we wish to contrast. Sjölander (2021) showed how this marginal effect can be estimated with a fixed effects model. The estimator relies on the fact that, if there are no unmeasured nonshared confounders, the regression function $\alpha_i + \gamma C_{ij}^m + \beta x$ can be viewed as a prediction of the potential outcome $Y_{ij}(x)$. Hence, for any fixed value x the mean potential outcome $\mathrm{E}\{Y_{ij}(x)\}$ can be expressed as

$$\mathrm{E}\{Y_{ij}(x)\} = \mathrm{E}(\alpha_i + \gamma C_{ij}^m + \beta x), \quad 13.$$

where the expectation is taken over (α_i, C_{ij}^m). When the link function g is the identity link or the log link, Sjölander (2021) showed that it is possible to construct an estimate of α_i, for each family i, such that $g(\alpha_i)$ is unbiased. Plugging these estimates into Equation 13 together with the conditional ML estimates of (γ, β), and replacing the expectation with the sample average, gives a consistent estimate of $\mathrm{E}\{Y_{ij}(x)\}$. This estimation method is referred to as standardization.

Unfortunately, there is currently no analogous estimator of $\mathrm{E}\{Y_{ij}(x)\}$ when g is the logit link. An alternative standardization approach, which works for any link function, is based on the BW model (Brumback et al. 2010, Cai & Brumback 2015). This approach uses the fact that, under the additional assumption in Equation 4, the predictor of $Y_{ij}(x)$ can be written as $\alpha_i^\dagger + \gamma_B \overline{C}_{i\cdot}^m + \beta_B \overline{X}_{i\cdot} + \gamma C_{ij}^m + \beta x$. Hence, for any fixed value x, the mean potential outcome $\mathrm{E}\{Y_{ij}(x)\}$ can be written as

$$\mathrm{E}\{Y_{ij}(x)\} = \mathrm{E}(\alpha_i^\dagger + \gamma_B \overline{C}_{i\cdot}^m + \beta_B \overline{X}_{i\cdot} + \gamma C_{ij}^m + \beta x), \quad 14.$$

where the expectation is taken over $(\alpha_i^\dagger, \overline{C}_{i\cdot}^m, \overline{X}_{i\cdot}, C_{ij}^m)$. A consistent estimate of $\mathrm{E}\{Y_{ij}(x)\}$ is obtained by replacing $(\gamma_B, \beta_B, \gamma, \beta)$ with their estimates, $\alpha_i^\dagger$ with a model-based prediction for each family i, and the expectation with the sample average (Brumback et al. 2010). Alternatively, one may integrate out $\alpha_i^\dagger$ from the expectation by using its estimated marginal distribution (Cai & Brumback 2015). A similar estimator for time-to-event outcomes was proposed by Dahlqwist et al. (2019).

The estimator based on the BW model in Equation 5 is computationally demanding and may not be feasible for large data sets. However, there is an alternative BW model for which estimation is easier. In this model the random intercept α_i is replaced with a fixed intercept α, common for

all families. The model is thus given by

$$g\{\mathrm{E}(Y_{ij}|\mathbf{C}_i^m, \mathbf{X}_i)\} = \alpha + \gamma_B \overline{C}_{i\cdot}^m + \beta_B \overline{X}_{i\cdot} + \gamma C_{ij}^m + \beta X_{ij}. \quad 15.$$

To distinguish between the BW models in Equations 5 and 15, Sjölander (2021) referred to the former as conditional and the latter as marginal. The marginal BW model appears to be more common than the conditional BW model in certain fields, particularly in twin research (Carlin et al. 2005). Sjölander et al. (2012a) discussed the interpretation of the parameter β in the marginal model. They showed that β may be interpreted as being partly marginal and partly conditional; it is marginal over the shared confounders C_i but conditional over the measured nonshared confounders C_{ij}^m. Sjölander (2021) showed how the marginal BW model can be used to estimate the fully marginal effect in Equation 12 by averaging over the measured nonshared confounders. Since the marginal BW model is a standard generalized linear model that does not require numerical integration of complex likelihoods, this estimator can be computed very quickly with standard software, even for large data sets.

3.3.2. Exposure regression. Skinner & D'Arragio (2011) showed how marginal causal effects can be estimated with the exposure regression in Equation 7, provided that there are no unmeasured nonshared confounders, the exposure is binary, and g^* is the logit link. Their approach uses inverse probability weighting (IPW) and is similar to the estimation of causal effects in marginal structural models (Hernán & Robins 2020). However, an important difference is that the model in Equation 7 has a family-specific intercept α_i^*, which complicates the weighting scheme. In line with standard conditional logistic regression, Skinner & D'Arragio (2011) proposed to condition on the sufficient statistic $\sum_j X_{ij}$ for each family i, which eliminates α_i from the weights.

While avoiding the incidental parameter problem, this conditioning induces another problem—namely, that some siblings will have infinitely large weights. This happens when one attempts to estimate the mean potential outcome $\mathrm{E}\{Y_{ij}(x)\}$ for a specific exposure level x (0 or 1), and there are families where no sibling has this exposure level. For such siblings, the conditional probability of receiving $X_{ij} = x$, given $\sum_j X_{ij}$, is 0, which gives infinity when inversely weighted with. A simple solution to this problem is to restrict the analysis to the exposure-discordant siblings, i.e., to the families with at least one exposed and one unexposed sibling, thereby estimating the marginal causal effect in this subsample (Sjölander 2021). In the special case when all families have exactly two siblings, this becomes the marginal causal effect among the exposure-discordant pairs (Equation 11). Current open questions are whether exposure regression can be used to estimate the marginal causal effect in the whole sample (Equation 12) and whether it can accommodate nonbinary exposures.

3.4. Marginal or Conditional Causal Effects?

Marginal and conditional causal effects answer different research questions, and the choice between these should ideally be driven by subject matter interest and scientific relevance. For instance, if the aim is to estimate the effect of imposing an intervention (e.g., preventing an exposure) from the whole population, then the marginal effect in Equation 12 is a relevant target parameter. If the intervention may only be imposed on a subgroup of the population, then the conditional effect in Equation 3 may more accurately reflect its effect.

As noted earlier, we may sometimes wish to make the models in Equations 1 and 5 more realistic by adding, for instance, splines or interactions between X_{ij} and C_{ij}^m. This generally makes the interpretation of conditional effects more difficult, since the conditional exposure-outcome

relation is then quantified by several parameters simultaneously. For instance, when C_{ij}^{m} contains five measured confounders, there are also five possible exposure-confounder interactions that we may wish to include in the model, which are all part of the conditional effect. However, the interpretation of marginal effects does not necessarily become more difficult since these average out over all interactions. Hence, by focusing on marginal causal effects, the researcher is relieved from the pressure to keep interpretation simple by using overly simplistic models.

That said, we note that the aim of some studies is precisely to detect interactions between the exposure and predictors for the outcome (e.g., confounders). For instance, when the aim is to implement an intervention, one may want to understand how the effect varies across subgroups (e.g., who is helped by the new treatment). Even when the aim is not to implement interventions, but rather to understand the etiology of the exposure-outcome relation, it may often be important to investigate how this relation varies across subgroups. If so, then marginal effects are irrelevant, since these are uninformative about effect heterogeneity.

Finally, we note that there are situations in which marginal and conditional effects coincide. In the models in Equations 1 and 5, we have assumed that the conditional effect is constant ($= \beta$) across levels of the confounders. Under this assumption, and if g is the identity link or log link, the marginal causal effect in Equation 12, with $h = g$ and $x' = x + 1$, is equal to the conditional causal effect in Equation 3. This is intuitively reasonable; if the exposure effect is the same for, say, women and men, we would expect that it is also the same in the mixed sample of women and men. Hence, in this special case, nothing is gained by standardization. However, for other link functions such as the logit link, marginal and conditional effects are not generally equal, even when the conditional effect is constant across levels of the confounders; this phenomenon is referred to as noncollapsibility (Greenland et al. 1999). In the special case when there is no confounding, and the family-specific intercept α_i has a normal distribution with variance σ^2, it can be shown that the conditional odds ratio is approximately $\sqrt{1 + 0.346\sigma^2}$ times the marginal odds ratio (Fitzmaurice et al. 2011). This indicates that the degree of noncollapsibility tends to increase with the heterogeneity in the outcome over clusters, e.g., with σ^2.

3.5. Software

Most of the methods and models that we have reviewed have been implemented in standard software, and new implementations are frequently made. We provide here a few examples that we have experience with, which appear robust and well programmed.

In R, conditional logistic regressions and stratified Cox regressions can be fitted with the `clogit` and `coxph` functions, respectively, in the `survival` package. The fixed effects model in Equation 1 can be fitted with the `gee` function in the `drgee` package; this function uses the semiparametric estimator of Goetgeluk & Vansteelandt (2008). The BW model in Equation 5 can be fitted with functions for random effects models, such as the `glmer` function in the `lme4` package. Estimation of conditional effects with the exposure regression in Equation 7, as well as DR estimation, can be carried out with the `drgee` function. Estimation of marginal causal effects with (log-)linear fixed effects models or marginal BW models can be carried out with the `stdGee` and `stdGlm` functions, respectively, in the `stdReg` package.

In Stata, conditional logistic regressions and stratified Cox regressions can be fitted with the `clogit` and `stcox` commands, respectively. The fixed effects model in Equation 1 and the BW model in Equation 5 with identity, log, and logit links can be fitted with the `xtreg`, `xtpoisson`, and `xtlogit` commands, respectively; for fixed effects models these functions use the conditional ML estimator. Estimation of marginal causal effects with marginal BW models can be carried out with the `margins` command.

In SAS, conditional logistic regressions and stratified Cox regressions can be fitted with the `LOGISTIC` and `PHREG` procedures, respectively. We have found no general implementation of the conditional ML estimator or the semiparametric estimator of Goetgeluk & Vansteelandt (2008) in SAS, but Allison (2005) provides workarounds for several link functions of the fixed effects model in Equation 1. The BW model in Equation 5 can be fitted with the `GLIMMIX` procedure.

4. METHODOLOGICAL CHALLENGES

Although sibling comparison studies are powerful tools to adjust for unmeasured confounding, they have special challenges that are not present, or present to a lesser extent, in studies of unrelated individuals. Several of these challenges arise because in order to estimate the exposure effect, sibling comparison studies mainly use information from the exposure-discordant siblings and largely ignore the exposure-concordant siblings. In particular, sibling comparison studies ignore all subjects without siblings. This is obvious for the model-free analysis of exposure-discordant pairs and the IPW analysis of Skinner & D'Arragio (2011), but it is also true for the fixed effects model discussed in Section 3.2.1, regardless of whether one uses the conditional ML estimator or the semiparametric estimator of Goetgeluk & Vansteelandt (2008). On a superficial level, the BW model may seem to use information from all siblings; however, the close similarity between the estimates from the fixed effects model and the BW model indicates that the latter mainly relies on the same subsample of exposure-discordant siblings as the former. An important exception is the BW model for time-to-event outcomes, which seems to borrow information from all siblings, thereby producing a more efficient estimate than the stratified Cox regression (Sjölander et al. 2013). The standardization methods described in Section 3.3.1 use information from all siblings when averaging in the expressions in Equations 13 and 14; this is how they are able to estimate the marginal effect (Equation 12) in the whole sample. However, these averages are functions of the conditional effect β, which is estimated by mainly using information from the exposure-discordant siblings.

The distinction between informative and noninformative siblings is useful to convey the special challenges of sibling comparison studies, but this distinction is not always clear-cut. First, some of the exposure-concordant siblings may have a slight influence on the estimated exposure effect. This happens if there are measured nonshared confounders and the exposure-concordant siblings are discordant in these confounders. The siblings will then contribute to the estimated confounder effects, and since the exposure and confounder effects are not orthogonal (e.g., the likelihood for β and γ under the models in Equation 1 or 5 does not factorize), the siblings may therefore also contribute indirectly to the estimated exposure effect. Second, when the exposure is truly continuous and measured with high accuracy, there will be no siblings that are perfectly exposure concordant since no individuals will have exactly the same exposure levels. In practice, though, one would typically expect that families with little or no variation in the exposure contribute relatively little to the estimated exposure effect. For pedagogical purposes, we thus argue in the following sections as if the distinction between informative and noninformative siblings is clear-cut and as if sibling comparison studies are entirely restricted to exposure-discordant siblings.

4.1. Low Statistical Power/Efficiency

An obvious consequence of the restriction to exposure-discordant siblings is a loss of statistical power/efficiency. This becomes clear when contrasting the sibling comparison estimate with a population-level estimate obtained by treating the siblings as unrelated individuals, e.g., by analyzing the siblings with ordinary linear or logistic regression. Whereas the population-level estimate is often expected to have larger bias, since it is not adjusted for any unmeasured confounders

shared within families, the sibling comparison estimate often has much higher variance—e.g., wider confidence intervals—due to the smaller effective sample size.

One way to handle this problem is to weigh the two estimates together to minimize mean squared error (MSE) (Kalish 1990, Greenland 1991, Sjölander 2013). However, the optimal weights that are guaranteed to minimize the MSE are usually unknown, and there is no guarantee that the MSE is minimized in finite samples when the weights are estimated from data. Furthermore, optimal weights have currently only been derived for odds ratios, and without taking measured confounders into account. Finally, due to noncollapsibility, the two estimated odds ratios will generally have different interpretations, even in the complete absence of confounding. Hence, it is not straightforward to interpret the (asymptotic limit of the) weighed estimate.

4.2. Poor Generalizability to the Whole Sample

Another consequence of the restriction to exposure-discordant siblings is that the sibling comparison estimates may not generalize well to the whole sample. Both the fixed effects model in Equation 1 and the BW model in Equation 5 assume that the exposure effect is constant and equal to β across all levels of C_i, i.e., across all families. Under this assumption, the effect that is estimated from the exposure-discordant siblings applies to all siblings. This is a crucial assumption for the standardization methods discussed in Section 3.3.1, and it is the handle that allows these methods to estimate the marginal effect (Equation 12) in the whole sample. However, the assumption of a constant exposure effect is strong and untestable, and it is likely violated to some extent in any real scenario. When the assumption is violated, the estimate of β is a weighted average of the C_i-specific effects, where the weight depends on the distribution of C_i among the exposure-discordant siblings (Sjölander et al. 2012b, Sjölander & Zetterqvist 2017). One may thus argue that, in practice, the analysis based on the fixed effects/BW model may not be more informative about the whole sample of both exposure-discordant and exposure-concordant siblings than is the model-free analysis or the IPW analysis by Skinner & D'Arragio (2011). The difference, one may argue, is that the latter analyses are more honest about what can reasonably be learned from data, in that they refrain from making strong extrapolations from exposure-discordant to exposure-concordant siblings and are more explicit about what parameter is being estimated in practice—i.e., the marginal effect (Equation 11) among the exposure-discordant siblings.

One way to assess the degree of generalizability from sibling comparison studies is to compare the distribution of measured characteristics between exposure-discordant and exposure-concordant siblings. If these differ substantially, then this indicates that the siblings that are informative for sibling comparisons are special in some sense and that the sibling comparison estimates may not generalize well to the whole sample. However, we note that this method is at best indicative, since it does not take unmeasured confounders into account.

4.3. Amplified Bias Due to Unmeasured Nonshared Confounding

In most observational studies, there are unmeasured confounders. When the unmeasured confounders are not shared within families, it can be shown that sibling comparison studies often tend to amplify the bias due to these, as compared with studies of unrelated individuals (Griliches 1979, Frisell et al. 2012). This happens when the exposure is correlated within families due to other reasons than shared confounding, as illustrated by the double dashed arrow between X_{i1} and X_{i2} in the causal diagram of **Figure 2** for a family with two siblings. Such correlation is likely present, to some extent, in virtually all sibling comparison studies. In **Figure 2** we absorb the shared confounders C_i into the nonshared confounders C_{ij}, thereby making these partly shared, and thus also

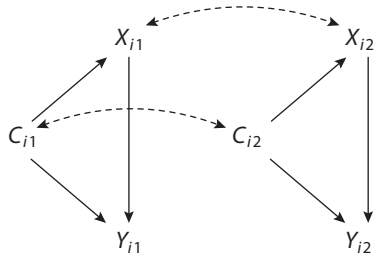

Figure 2

Causal diagram illustrating amplified bias due to unmeasured nonshared confounding (the *dashed double arrow* between X_{i1} and X_{i2}), for a family with two siblings. The shared confounders C_i have been absorbed into the nonshared confounders C_{ij}, making them partly shared and thus also correlated, as illustrated by the dashed double arrow between C_{i1} and C_{i2}.

correlated. This latter modification is mainly for pedagogical purposes, as it gives more intuitive bias expressions. However, one may also argue that the clear-cut distinction between shared and nonshared confounders in **Figure 1** is somewhat artificial, since very few confounders are in practice exactly equal within families. For instance, even monozygotic twins may have slightly different genomes due to postfertilization mutations.

To understand why the bias is amplified, suppose for pedagogical purposes that all variables are binary and that all arrows on the diagram represent positive effects—e.g., $C_{ij} = 1$ increases the probability of $X_{ij} = 1$. Due to the correlation between X_{i1} and X_{i2}, these exposures tend to be equal in a random pair taken from the whole sample. However, sibling comparison studies are restricted to the exposure-discordant pairs, i.e., to the subsample where X_{i1} and X_{i2} are not equal. In order for X_{i1} and X_{i2} to differ, there has to be a differential influence of other factors, presumably C_{i1} and C_{i2}. Hence, to explain that, say, $X_{i1} = 0$ and $X_{i2} = 1$, we would expect that $C_{i1} = 0$ and $C_{i2} = 1$ as well—i.e., we would expect a correlation between C_{ij} and X_{ij}. Of course, even in the whole sample of all pairs, there is a correlation between C_{ij} and X_{ij} due to the positive effect of C_{ij} on X_{ij}, but in the subsample of exposure-discordant pairs, this correlation has to be stronger in order to override the correlation between X_{i1} and X_{i2} and make these unequal. As the selection imposes a stronger correlation between C_{ij} and X_{ij}, the confounding influence by C_{ij} becomes stronger as well.

Griliches (1979) and Frisell et al. (2012) derived analytic bias formulas under linear models, and Frisell et al. (2012) provided simulations under logistic models for the scenario in **Figure 2**. The formulas for linear models are pedagogically useful, so we reproduce one of them here. Suppose that C_{ij} in **Figure 2** is completely unmeasured, and let $bias_{\text{fixed}}$ and $bias_{\text{lm}}$ be the biases of the sibling comparison estimate and the population-level estimate of β, obtained from a linear fixed effects model and an ordinary linear regression model, respectively. The ratio of biases is given by

$$\frac{bias_{\text{fixed}}}{bias_{\text{lm}}} = \frac{\gamma^2\sigma_C^2 + \sigma_X^2}{\gamma^2\sigma_C^2 + \sigma_X^2 \frac{1-\rho_X}{1-\rho_C}}. \qquad 16.$$

In this expression, γ is the effect of C_{ij} on X_{ij}; σ_C^2 is the variance of C_{ij}; ρ_C is the correlation between C_{i1} and C_{i2}; σ_X^2 is the conditional (residual) variance of X_{ij}, given C_{ij}; and ρ_X is the conditional correlation between X_{i1} and X_{i2}, given C_{i1} and C_{i2}. Thus, ρ_C parameterizes the dashed double arrow between C_{i1} and C_{i2} in **Figure 2**, and ρ_X parameterizes the dashed double arrow between X_{i1} and X_{i2}. The numerator and denominator in Equation 16 differ by the term $\frac{1-\rho_X}{1-\rho_C}$. When ρ_C increases, this term increases as well, so that the bias ratio decreases. This makes intuitive

sense since, in the limit when C_{ij} is perfectly shared within families ($\rho_C = 1$), the fixed effects model eliminates all confounding bias so that the bias ratio is 0. However, when ρ_X increases, the bias ratio increases as well. This is a consequence of the restriction to exposure-discordant pairs explained above; the larger the correlation ρ_X, the stronger the confounding influence by C_{ij} in this restricted subsample. When $\rho_X > \rho_C$, the fixed effects model gives larger bias than the ordinary linear regression, and thus the elimination of bias due to shared confounders is outweighed by the amplified bias due to unmeasured nonshared confounders.

4.4. Attenuated Effect Due to Measurement Errors in the Exposure

In many studies, the observed variables are measured with error. It can be shown that random measurement errors in the exposure tend to attenuate the estimated effects more strongly in sibling comparison studies than in studies of unrelated individuals (Griliches 1979, Frisell et al. 2012). This happens for similar reasons as the bias amplification discussed in the previous section. Let X^*_{ij} be the observed exposure level for sibling j within family i; in the presence of measurement error, X^*_{ij} is not necessarily equal to X_{ij}. The subsample of apparently exposure-discordant pairs ($X^*_{i1} \neq X^*_{i2}$) is a mixture of pairs that are truly exposure discordant and pairs that are truly exposure concordant. If the correlation ρ_X is large, then the true exposures X_{i1} and X_{i2} tend to be equal in any given pair. Hence, if the observed exposures X^*_{i1} and X^*_{i2} are not equal, then we may suspect that this is due to measurement error and that the apparently exposure-discordant pair is truly exposure concordant. The larger the correlation ρ_X, the more likely this misclassification is to happen, and the greater the attenuation of the estimated effect in the selected subsample of apparently exposure-discordant pairs will be.

Ashenfelter & Krueger (1994) proposed a way to correct for this measurement error bias when more than one measure of the exposure is available. In their co-twin control study of economic returns to schooling, each twin was asked to provide information about her co-twin's education history as well her own. By having two measures of the exposure for each twin and assuming an additive model for the measurement errors with uncorrelated errors, the authors were able to derive a bias-corrected estimate for the fixed effects model with identity link. However, there is currently no extension of the method for models with other link functions.

4.5. Bias Due to Carryover Effects

The causal diagram in **Figure 1** assumes that the exposure and outcome of one sibling have no effect on the exposure and/or outcome of the other sibling. However, such carryover effects may often be present in real studies. Sjölander et al. (2016) provided a discussion of various types of carryover effects and their implications. They showed that carryover effects generally lead to biased estimates in sibling comparison studies, and they derived bias expressions under fixed effects models with identity and logit links. They concluded that the bias tends to attenuate the estimated effect toward the null in some common scenarios, thus producing a conservative estimate of the true exposure effect. Furthermore, some types of carryover effects may not give bias under the null hypothesis of no exposure effect, in which case the statistical test of this null hypothesis remains valid.

As an example of how carryover effects may give bias, consider the sibling comparison study of Meyer et al. (2004), where the exposure was maternal smoking and the outcome was birth defects in the offspring. Suppose that women who give birth to a child with birth defects tend to avoid smoking during the next pregnancy to minimize the risk that birth defects occur in the subsequent child as well. This outcome-to-exposure carryover effect is represented in the causal diagram of

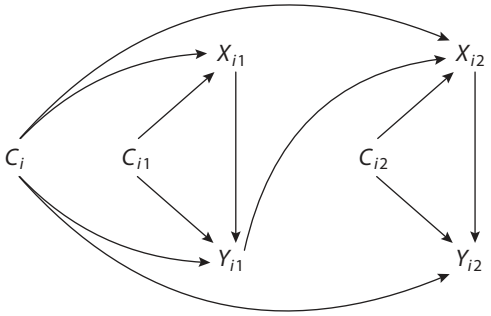

Figure 3

Causal diagram illustrating exposure-to-outcome carryover effects, for a family with two siblings. The arrow from Y_{i1} to X_{i2} represents an outcome-to-exposure carryover effect.

Figure 3. A consequence of this carryover effect is that children who are unexposed are more likely to have an older sibling with birth defects than children who are exposed. Thus, among the exposure-discordant siblings (e.g., first sibling exposed, second sibling unexposed), the outcome tends to be more common among the exposed sibling than among the unexposed sibling, in the complete absence of an exposure effect. Hence, for this type of carryover effect, even a statistical test of the null hypothesis is invalid. Sjölander et al. (2016) gave a more general explanation for the bias induced by carryover effects, based on properties of conditional ML estimators. Petersen & Lange (2020) provided further discussion of the symmetric scenario where there is a carryover effect of X_{i1} on Y_{i2} and a carryover effect of X_{i2} on Y_{i1}. They showed that, although the conditional ML estimate of β in the fixed effects model is biased under this scenario, it still has a causal interpretation in the absence of unmeasured nonshared confounders. However, for nonlinear models, this causal interpretation is rather nonstandard as it simultaneously includes the exposures and outcomes for both siblings.

One may attempt to use the observed data to test for the presence of carryover effects. For instance, Meyer et al. (2004) regressed the exposure X_{i2} of the second sibling on the outcome Y_{i1} of the first sibling, while controlling for the exposure X_{i1} of the first sibling. When observing no conditional association between Y_{i1} and X_{i2}, these authors concluded that birth defects in the first sibling are not likely to influence maternal smoking behavior in the second sibling. The authors' rationale for this analysis can be understood from the causal diagram in **Figure 3**; conditioning on X_{i1} blocks the path $Y_{i1} \leftarrow X_{i1} \leftarrow C_i \rightarrow X_{i2}$, and thus one would perhaps expect Y_{i1} and X_{i2} to be conditionally independent, unless there was also a path $Y_{i1} \rightarrow X_{i2}$. However, the causal diagram also reveals that the conditioning on X_{i1} opens the path $Y_{i1} \leftarrow C_{i1} \rightarrow X_{i1} \leftarrow C_i \rightarrow X_{i2}$ on which X_{i1} acts as a collider (Greenland 2003). Hence, the absence of conditional association between Y_{i1} and X_{i2} may, at least in principle, also be explained by an association component along this opened path that has the opposite sign of, and almost perfectly cancels with, a carryover effect of Y_{i1} on X_{i2}. This example illustrates that carryover effects are difficult to test for, due to the (at least partly) unmeasured confounders C_{ij} and C_i. Sjölander et al. (2016) provided other examples as well as possible strategies to reduce bias for some types of carryover effects.

4.6. Shared Mediators

We have argued that sibling comparison studies automatically adjust for all shared confounders. However, covariates that are shared within the family may not necessarily be confounders, which

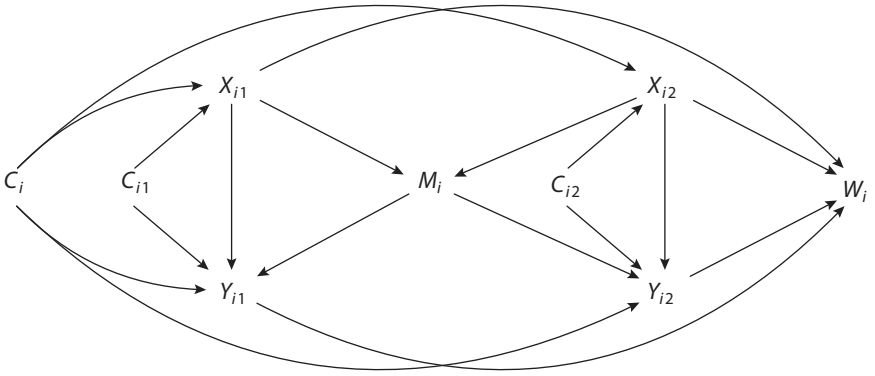

Figure 4

Causal diagram illustrating shared confounders (C_i), mediators (M_i), and colliders (W_i), for a family with two siblings.

is illustrated by the causal diagram of **Figure 4** for a family with two siblings. In this diagram, the variable M_i represents the set of shared mediators; these are affected by the exposures (X_{i1}, X_{i2}) and in turn affect the outcomes (Y_{i1}, Y_{i2}). The variable W_i represents the set of shared colliders; these are affected by both the exposures (X_{i1}, X_{i2}) and the outcomes (Y_{i1}, Y_{i2}). Since the fixed effects model and the BW model adjust for the shared confounders in an implicit fashion, by absorbing these into the family-specific intercept α_i, one may wonder whether these models also absorb and thereby implicitly adjust for the shared mediators and colliders. Using both analytic arguments and simulations, Sjölander & Zetterqvist (2017) showed that the models do indeed adjust for shared mediators as well as shared confounders, but not for shared colliders.

It follows that the estimated effect in a sibling comparison study may not be interpreted as a total exposure effect, but rather as a direct effect from which all influence through shared mediators has been washed out. We emphasize that this may not necessarily be viewed as a bias, but as an inherent feature of sibling comparison studies to implicitly define the target parameter as a direct effect. Whether this feature is positive or not depends on whether the direct or total effect is most relevant for the research question at hand. That the estimate is not adjusted for shared colliders is entirely positive though, since adjustment for colliders would lead to bias—e.g., to an observed exposure-outcome association even in the complete absence of a causal effect (Greenland 2003).

5. ILLUSTRATION: FETAL GROWTH RESTRICTION AND ATTENTION-DEFICIT/HYPERACTIVITY DISORDER

Several studies have shown that poor fetal growth is statistically associated with various neuropsychiatric conditions, such as autism spectrum disorder (Lampi et al. 2012), depression and bipolar disorder (Nosarti et al. 2012), and schizophrenia (Abel et al. 2010). However, the strong potential for confounding by unmeasured (e.g., genetic) factors raises concerns about whether these statistical associations represent causal relations. To adjust for unmeasured familial confounding in this context, Pettersson et al. (2015) carried out a co-twin control study of the association between fetal growth restriction and ADHD symptoms. Using data from the Swedish Twin Registry, they found a statistically significant association between low (for gestational age) birth weight and high degree of ADHD symptoms. As an illustration, we elaborate on and extend the analysis by Pettersson et al. (2015).

5.1. Data

Birth weight was obtained from the Medical Birth Registry, which contains data from more than 95% of all births in Sweden. ADHD symptoms were assessed when the twins were between 9 and 12 years old, through telephone interviews with their parents. The symptoms were measured with two continuous scores, ranging from 0 to 9, corresponding to inattention and hyperactivity–impulsivity symptoms, respectively. Finally, these two scores were added to produce a total ADHD score ranging from 0 to 18. Pettersson et al. (2015) included both monozygotic and dizygotic (DZ) twins in their analysis; we focus here on the DZ twins. After excluding pairs with missing information, the original data set comprises 11,816 twins from 5,908 pairs, born between 1992 and 2000. These data are not publicly available. Thus, to enable the reader to replicate our analyses, we simulated data that are very similar, but not identical, to the real data. R code for the simulation is provided in the **Supplemental Appendix**.

Supplemental Material >

Pettersson et al. (2015) considered birth weight as their exposure and adjusted for gestational age in all analyses. Arguably, though, the underlying causal agent (if any) is instead fetal growth, which is captured by birth weight and gestational age jointly. Thus, we defined our exposure as the standardized birth weight,

$$birth\ weight_{\text{std}} = \frac{birth\ weight - \text{E}(birth\ weight|gestational\ age)}{\text{sd}(birth\ weight|gestational\ age)},$$

where E(*birth weight*|*gestational age*) and sd(*birth weight*|*gestational age*) are the gestational age–specific mean and standard deviation, respectively (Land 2006). In this definition we used the means and standard deviations in the general population of all infants as estimated by Maršál et al. (1996), not the mean and standard deviation in the subpopulation of twins. We also used a binary version of the exposure, where we followed a common convention and defined infants with standardized birth weight below the 10th percentile (i.e., $birth\ weight_{\text{std}} < -1.28$, assuming that $birth\ weight_{\text{std}}$ has a standard normal distribution) as small for gestational age (SGA) (Wikipedia 2020).

5.2. Descriptive Statistics

Figure 5 shows the distribution of standardized birth weight (panel *a*) and ADHD score (panel *b*) for the DZ twins. On average, the twins are considerably smaller at birth than infants from the general population; the mean standardized birth weight is −0.66. Most twins have few or no ADHD symptoms; the mean ADHD score is 1.55, and 54% of all twins have a score equal to 0. The within twin-pair correlations for standardized birth weight and ADHD score are 0.35 and 0.25, respectively. Among all twin pairs, 32.3% are discordant in SGA.

5.3. Methods

We first carried out population-level analyses, which do not adjust for any unmeasured confounders. In these analyses we estimated the mean difference in ADHD score between the SGA and non-SGA twins, and fitted an ordinary linear regression model with standardized birth weight as the exposure and the ADHD score as the outcome, adjusted for infant sex. We computed 95% confidence intervals for the mean difference and the estimated regression slope, using the sandwich formula (Stefanski & Boos 2002) to account for the paired data structure.

To adjust for unmeasured confounders that are shared within pairs of DZ twins, we carried out co-twin control comparisons. We again estimated the mean difference in ADHD score

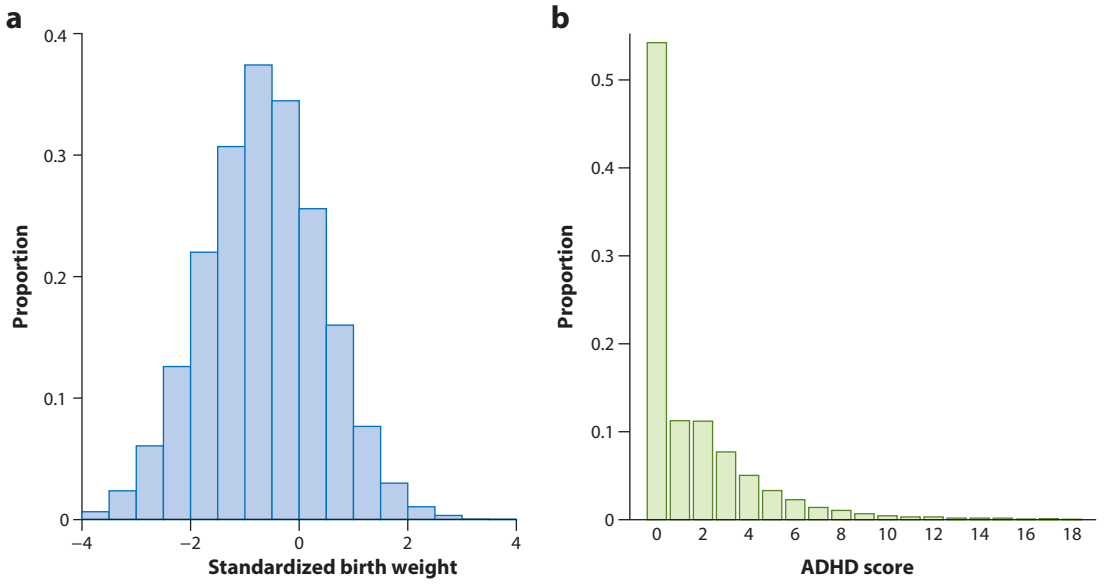

Figure 5

Distribution of standardized birth weight and ADHD score for dizygotic twins from the Swedish Medical Birth Registry. Abbreviation: ADHD, attention-deficit/hyperactivity disorder.

between the SGA and non-SGA twins, now restricting the analysis to the SGA-discordant twin pairs. We fitted a linear fixed effects model and a linear BW model, with standardized birth weight as the exposure and the ADHD score as the outcome, adjusted for infant sex. As a comparison, we also estimated the exposure coefficient β in these models with exposure regression, using a linear fixed effects model for standardized birth weight adjusted for infant sex. Finally, to allow for nonlinear effects, we refitted the fixed effects model for the outcome, replacing the linear exposure term by a natural cubic spline function with knots at the three quartiles in the standardized birth weight distribution (−1.40, −0.65, and −0.06) of the full sample. To allow for sex-specific effects, we added an interaction (product) term between this spline function and the sex variable in the model. Based on the fitted model, we used regression standardization to estimate the mean counterfactual ADHD score as a function of standardized birth weight. We provide R code for all analyses in the **Supplemental Appendix**.

Supplemental Material >

5.4. Results

Table 1 shows the estimated mean difference between SGA and non-SGA twins and the estimated exposure coefficient β together with 95% (Wald) confidence intervals, for the population-level analyses and the co-twin control comparisons. The population-level analyses indicate that SGA infants have 0.13 units higher ADHD score than non-SGA infants, on average, and that 1 unit increase in standardized birth weight is associated with a 0.11 unit decrease in ADHD score, on average. These associations are somewhat stronger in the co-twin control comparisons, where the corresponding figures are 0.21 and −0.22, respectively. The estimates and confidence intervals are virtually identical for the two outcome regressions and the exposure regression, which is expected from theory (Section 3.2.1). All associations are statistically significant, at 5% significance level. The estimated between-effect β_B in the BW model (not shown in **Table 1**) was equal to 0.12,

Table 1 Analysis results

	Estimate	95% CI
Population-level analysis		
SGA versus non-SGA	0.11[a]	(0.00, 0.21)
Ordinary linear regression	−0.11[b]	(−0.16, −0.07)
Co-twin control		
SGA versus non-SGA	0.22[a]	(0.08, 0.36)
Outcome regression, fixed effects model	−0.19[b]	(−0.26, −0.13)
Outcome regression, BW model	−0.19[b]	(−0.26, −0.13)
Exposure regression	−0.19[b]	(−0.26, −0.13)

[a]Estimated mean difference.
[b]Estimated exposure coefficient β, adjusted for sex.
Abbreviations: BW, between-within; CI, confidence interval; SGA, small for gestational age.

with a 95% confidence interval equal to (0.03, 0.20). The statistical significance of this estimate indicates fairly strong evidence of shared confounding.

Figure 6 shows the estimated mean counterfactual ADHD score as a function of standardized birth weight and the estimated mean difference, using 0 as reference, together with pointwise 95% confidence limits. We observe that the estimated effect indeed appears nonlinear and flattens out at higher standardized birth weights. At a standardized birth weight of 0, the estimated mean is ∼1.39, which indicates that, if all children would have a standardized birth weight equal to the general population mean, then the mean ADHD score would be about 1.39 units. We emphasize that this causal interpretation crucially hinges on the assumption of no nonshared confounders, except sex.

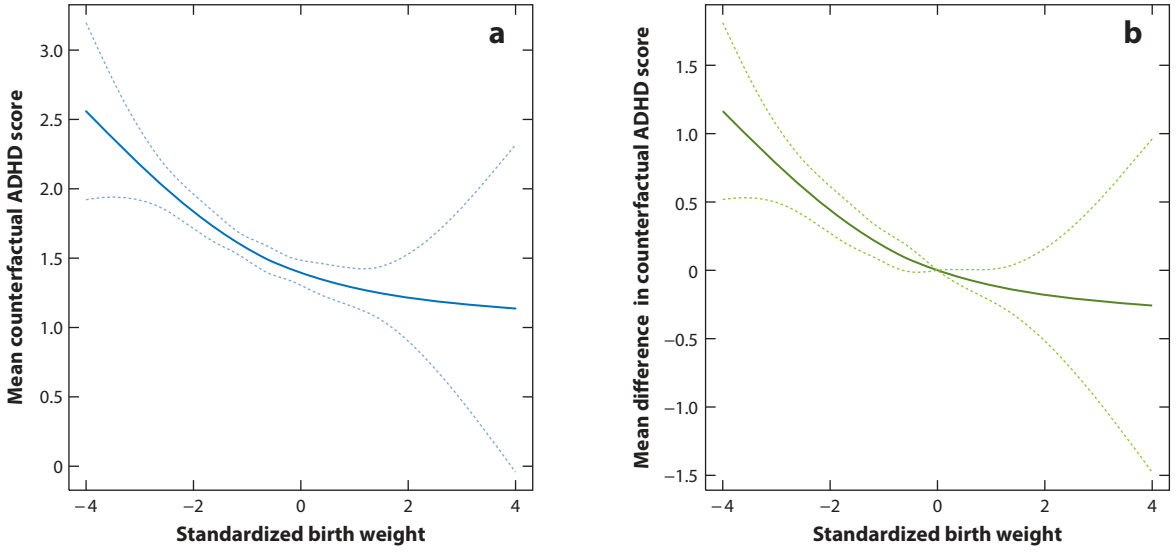

Figure 6

(*a*) Estimated mean counterfactual ADHD score as a function of standardized birth weight (*solid line*), with pointwise 95% confidence limits (*dashed lines*). (*b*) Mean difference in estimated counterfactual ADHD score as a function of standardized birth weight (*solid line*), using 0 as reference, with pointwise 95% confidence limits (*dashed lines*). Abbreviation: ADHD, attention-deficit/hyperactivity disorder.

5.5. Critical Interpretation of the Findings

The co-twin control comparisons provide fairly strong evidence for a causal effect of fetal growth restriction on ADHD symptoms. However, the results should be interpreted in light of the challenges discussed in Section 4. The loss of statistical power/efficiency and the potential for poor generalizability are strong concerns in the model-free analysis, since the effective sample size is reduced by two-thirds when dichotomizing birth weight. This may be less of a problem in the model-based analyses, which avoid dichotomization of birth weight and thereby enable information to be drawn from the whole sample.

The potential for amplified bias due to unmeasured nonshared confounding is a concern, since the standardized birth weights of two twins from the same pair are quite strongly correlated. If the correlation is largely due to shared (e.g., parental) factors that are not related to the ADHD score, then this may substantially amplify any existing bias due to unmeasured nonshared confounding.

At first glance, influential bias due to measurement errors in the exposure may not seem likely, since birth weight is easy to measure with high accuracy and gestational age is identical for both twins within the pair. However, as argued above, the putative causal agent is fetal growth restriction, for which standardized birth weight only serves as a proxy, and it should in general be recognized that the degree of (bias due to) measurement errors is ultimately determined by how well our proxy measure correlates with the underlying true exposure. This general point concerns measurement error, which is random with respect to the twins, however, and it may be argued in the case of fetal growth that many of the factors that cause birth weight to be a poor proxy of fetal growth in the whole population are shared by twins (e.g., maternal stature). If measurement error is indeed shared by twins, the co-twin control will remove such measurement error along with other shared factors. Taken together, we believe that the accurate measure of birth weight and the potential for shared rather than random measurement errors make it unlikely that measurement error has substantially biased the within-twin estimate in our example.

Outcome-to-exposure carryover effects are logically impossible due to the temporal order of variables; the exposures for both twins occur before the outcomes of both twins. In principle, though, there could be other types of carryover effects, such as an exposure-to-exposure effect or an exposure-to-outcome effect. The former would occur if, for instance, the larger size of one twin inhibits the growth of the other twin. This may indeed be expected, since twins share a finite supply line and a larger share pooled to one twin may be at direct expense of the other. The latter would occur if, for instance, a twin born very small receives strong parental attention during the first years in life, and this induces attention-seeking and rebellious behavior in the co-twin.

Finally, one could imagine that a causal effect of fetal growth restriction on ADHD symptoms is partly mediated through factors that are shared within families. For instance, if either or both twins are severely growth restricted, then this may influence the psychosocial environment in the family, which in turn may influence the behavior of both twins. Such shared mediated effects are implicitly eliminated in all co-twin control comparisons.

6. SUMMARY

Sibling comparison studies have an indisputable role in observational research. They have a strong potential to reduce confounding bias, and they can be used to estimate both conditional and marginal causal effects, with a regression model for either the exposure or the outcome. These regression models are typically fixed effects models or BW models. Most of the analysis methods are implemented in standard software, which makes them easily accessible to practitioners. However, sibling comparison studies suffer from several methodological challenges, which are not present, or present to less extent, in studies of unrelated individuals. These include potentially

low statistical power and generalizability, amplified bias due to unmeasured nonshared confounding and attenuated effects due to measurement errors in the exposure, and bias due to carryover effects and shared mediators. Hence, the results from sibling comparison studies must be interpreted carefully with these challenges in mind.

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

A.S. gratefully acknowledges financial support from the Swedish Research Council.

LITERATURE CITED

Abel K, Wicks S, Susser E, Dalman C, Pedersen M, et al. 2010. Birth weight, schizophrenia, and adult mental disorder: Is risk confined to the smallest babies? *Arch. Gen. Psychiatry* 67(9):923–30

Allison P. 2005. *Fixed Effects Regression Methods for Longitudinal Data Using SAS*. Cary, NC: SAS Inst.

Allison P. 2009. *Fixed Effects Regression Models*. Thousand Oaks, CA: SAGE

Andersen E. 1970. Asymptotic properties of conditional maximum-likelihood estimators. *J. R. Stat. Soc. Ser. B* 32(2):283–301

Ashenfelter O, Krueger A. 1994. Estimates of the economic return to schooling from a new sample of twins. *Am. Econ. Rev.* 84(5):1157–73

Bang H, Robins J. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4):962–73

Boone-Heinonen J, Biel F, Marshall N, Snowden J. 2020. Maternal pre-pregnancy BMI and size at birth: race/ethnicity-stratified, within-family associations in over 500,000 siblings. *Ann. Epidemiol.* 46:49–56.e5

Breslow N, Day N. 1980. *Statistical Methods in Cancer Research*, Vol. I: *The Analysis of Case-Control Studies*. Lyon, Fr.: IARC/WHO

Brumback B, Dailey A, Brumback L, Livingston M, He Z. 2010. Adjusting for confounding by cluster using generalized linear mixed models. *Stat. Probab. Lett.* 80(21–22):1650–54

Cai Z, Brumback B. 2015. Model-based standardization to adjust for unmeasured cluster-level confounders with complex survey data. *Stat. Med.* 34(15):2368–80

Carlin J, Gurrin L, Sterne J, Morley R, Dwyer T. 2005. Regression models for twin studies: a critical review. *Int. J. Epidemiol.* 34(5):1089–99

Class Q, Rickert M, Larsson H, Lichtenstein P, D'Onofrio B. 2014. Fetal growth and psychiatric and socioeconomic problems: population-based sibling comparison. *Br. J. Psychiatry* 205(5):355–61

Dahlqwist E, Pawitan Y, Sjölander A. 2019. Regression standardization and attributable fraction estimation with between-within frailty models for clustered survival data. *Stat. Methods Med. Res.* 28(2):462–85

Dai J, Mukamal K, Krasnow R, Swan G, Reed T. 2015. Higher usual alcohol consumption was associated with a lower 41-y mortality risk from coronary artery disease in men independent of genetic and common environmental factors: the prospective NHLBI Twin Study. *Am. J. Clin. Nutr.* 102(1):31–39

Daley D, Jacobsen R, Lange A, Sørensen A, Walldorf J. 2019. The economic burden of adult attention deficit hyperactivity disorder: a sibling comparison cost analysis. *Eur. Psychiatry* 61:41–48

D'Onofrio B, Rickert M, Frans E, Kuja-Halkola R, Almqvist C, et al. 2014. Paternal age at childbearing and offspring psychiatric and academic morbidity. *JAMA Psychiatry* 71(4):432–38

Eisen S, Goldberg J, True W, Henderson W. 1991. A co-twin control study of the effects of the Vietnam War on the self-reported physical health of veterans. *Am. J. Epidemiol.* 134(1):49–58

Falconer DS, Mackay T. 1996. *Introduction to Quantitative Genetics*. New York: Pearson. 4th ed.

Fitzmaurice G, Laird N, Ware J. 2011. *Applied Longitudinal Analysis*. New York: Wiley

Floderus B, Cederlöf R, Friberg L. 1988. Smoking and mortality: a 21-year follow-up based on the Swedish twin registry. *Int. J. Epidemiol.* 17(2):332–40
Frisell T, Öberg S, Kuja-Halkola R, Sjölander A. 2012. Sibling comparison designs: bias from non-shared confounders and measurement error. *Epidemiology* 23(5):713–20
Gesell A. 1942. The method of co-twin control. *Science* 95(2470):446–48
Goetgeluk S, Vansteelandt S. 2008. Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics* 64(3):772–80
Gorseline D. 1932. *The effect of schooling upon income*. PhD Thesis, Indiana Univ., Bloomington
Greenland S. 1991. Reducing mean squared error in the analysis of stratified epidemiologic studies. *Biometrics* 47:773–76
Greenland S. 2003. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 14(3):300–6
Greenland S, Robins J, Pearl J. 1999. Confounding and collapsibility in causal inference. *Stat. Sci.* 14(1):29–46
Griliches Z. 1979. Sibling models and data in economics: beginnings of a survey. *J. Political Econ.* 87(5):S37–S64
Hernán M, Robins J. 2020. *Causal Inference: What If*. Boca Raton, FL: Chapman & Hall/CRC
Holt J, Prentice R. 1974. Survival analyses in twin studies and matched pair experiments. *Biometrika* 61(1):17–30
Jonsson F, Wolk A, Pedersen N, Lichtenstein P, Terry P, et al. 2003. Obesity and hormone-dependent tumors: cohort and co-twin control studies based on the Swedish Twin Registry. *Int. J. Cancer* 106(4):594–99
Jung S. 1999. Rank tests for matched survival data. *Lifetime Data Anal.* 5(1):67–79
Kalish L. 1990. Reducing mean squared error in the analysis of pair-matched case-control studies. *Biometrics* 46(2):493–99
Kaplan E, Meier P. 1958. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* 53(282):457–81
Kendler K, Karkowski L, Prescott C. 1999. Causal relationship between stressful life events and the onset of major depression. *Am. J. Psychiatry* 156(6):837–41
Klein J, Moeschberger M. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer. 2nd ed.
Kolk M, Barclay K. 2019. Cognitive ability and fertility among Swedish men born 1951–1967: evidence from military conscription registers. *Proc. R. Soc. B* 286:20190359
Lampi K, Lehtonen L, Tran P, Suominen A, Lehti V, et al. 2012. Risk of autism spectrum disorders in low birth weight and small for gestational age infants. *J. Pediatr.* 161(5):830–36
Lancaster T. 2000. The incidental parameter problem since 1948. *J. Econom.* 95(2):391–413
Land J. 2006. How should we report on perinatal outcome? *Hum. Reprod.* 21(10):2638–39
Lawlor D, Clark H, Smith G, Leon D. 2006. Intrauterine growth and intelligence within sibling pairs: findings from the Aberdeen children of the 1950s cohort. *Pediatrics* 117(5):e894–e902
Lawlor D, Mortensen L, Nybo Andersen A. 2011. Mechanisms underlying the associations of maternal age with adverse perinatal outcomes: a sibling study of 264,695 Danish women and their firstborn offspring. *Int. J. Epidemiol.* 40(5):1205–14
Levit S. 1935. Twin investigations in the U.S.S.R. *J. Personal.* 3(3):188–93
Little R, Rubin D. 2000. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu. Rev. Public Health* 21:121–45
Lown E, Goldsby R, Mertens A, Greenfield T, Bond J, et al. 2008. Alcohol consumption patterns and risk factors among childhood cancer survivors compared to siblings and general population peers. *Addiction* 103(7):1139–48
Lundström S, Forsman M, Larsson H, Kerekes N, Serlachius E, et al. 2014. Childhood neurodevelopmental disorders and violent criminality: a sibling control study. *J. Autism Dev. Disord.* 44(11):2707–16
Maršál K, Persson P, Larsen T, Lilja H, Selbing A, Sultan B. 1996. Intrauterine growth curves based on ultrasonically estimated foetal weights. *Acta Paediatr.* 85(7):843–48
Meyer K, Williams P, Hernandez-Diaz S, Cnattingius S. 2004. Smoking and the risk of oral clefts: exploring the impact of study designs. *Epidemiology* 15(6):671–78
Mundlak Y. 1978. Pooling of time-series and cross-section data. *Econometrica* 46(1):69–85

Murray C. 2002. IQ and income inequality in a sample of sibling pairs from advantaged family backgrounds. *Am. Econ. Rev.* 92(2):339–43
Neuhaus J, McCulloch C. 2006. Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *J. R. Stat. Soc. Ser. B* 68(5):859–72
Nilsen T, Knudsen G, Gervin K, Brandt I, Røysamb E, et al. 2013. The Norwegian twin registry from a public health perspective: a research update. *Twin Res. Hum. Genet.* 16(1):285–95
Nosarti C, Reichenberg A, Murray R, Cnattingius S, Lambe M, et al. 2012. Preterm birth and psychiatric disorders in young adult life. *Arch. Gen. Psychiatry* 69(6):610–17
Pearl J. 1995. Causal diagrams for empirical research. *Biometrika* 82(4):669–88
Pearl J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge Univ. Press. 2nd ed.
Petersen A, Lange T. 2020. What is the causal interpretation of sibling comparison designs? *Epidemiology* 31(1):75–81
Pettersson E, Sjölander A, Almqvist C, Anckarsäter H, D'Onofrio B, et al. 2015. Birth weight as an independent predictor of ADHD symptoms: a within-twin pair analysis. *J. Child Psychol. Psychiatry* 56(4):453–59
Piirtola M, Jelenkovic A, Latvala A, Sund R, Honda C, et al. 2018. Association of current and former smoking with body mass index: a study of smoking discordant twin pairs from 21 twin cohorts. *PLOS ONE* 13(7):e0200140
Rosenbaum P. 2015. How to see more in observational studies: some new quasi-experimental devices. *Annu. Rev. Stat. Appl.* 2:21–48
Rosenbaum P. 2020. Modern algorithms for matching in observational studies. *Annu. Rev. Stat. Appl.* 7:143–76
Rubin D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66(5):688–701
Seaman S, Pavlou M, Copas A. 2014. Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Stat. Med.* 33(30):5371–87
Sjölander A. 2013. Reducing mean squared error in the analysis of binary paired data. *Epidemiol. Methods* 2(1):33–47
Sjölander A. 2021. Estimation of marginal causal effects in the presence of confounding by cluster. *Biostatistics* 22(3):598–612
Sjölander A, Frisell T, Kuja-Halkola R, Öberg S, Zetterqvist J. 2016. Carryover effects in sibling comparison designs. *Epidemiology* 27(6):852–58
Sjölander A, Frisell T, Öberg S. 2012a. Causal interpretation of between-within models for twin research. *Epidemiol. Methods* 1(1):217–37
Sjölander A, Johansson A, Lundholm C, Altman D, Almqvist C, Pawitan Y. 2012b. Analysis of 1:1 matched cohort studies and twin studies, with binary exposures and binary outcomes. *Stat. Sci.* 27(3):395–411
Sjölander A, Lichtenstein P, Larsson H, Pawitan Y. 2013. Between–within models for survival analysis. *Stat. Med.* 32(18):3067–76
Sjölander A, Zetterqvist J. 2017. Confounders, mediators, or colliders. *Epidemiology* 28(4):540–47
Skinner C, D'Arragio J. 2011. Inverse probability weighting for clustered nonresponse. *Biometrika* 98(4):953–66
Skytthe A, Ohm Kyvik K, Vilstrup Holm N, Christensen K. 2011. The Danish twin registry. *Scand. J. Public Health* 39(7):75–78
Stefanski L, Boos D. 2002. The calculus of M-estimation. *Am. Stat.* 56(1):29–38
Sullivan W. 1899. A note on the influence of maternal inebriety on the offspring. *J. Mental Sci.* 45(190):489–503
Wikipedia. 2020. Small for gestational age. *Wikipedia.* **https://en.wikipedia.org/w/index.php?title=Small_for_gestational_age&oldid=995728915**
Zagai U, Lichtenstein P, Pedersen N, Magnusson P. 2019. The Swedish twin registry: content and management as a research infrastructure. *Twin Res. Hum. Genet.* 22(6):672–80
Zetterqvist J, Vansteelandt S, Pawitan Y, Sjölander A. 2016. Doubly robust methods for handling confounding by cluster. *Biostatistics* 17(2):264–76
Zetterqvist J, Vermeulen K, Vansteelandt S, Sjölander A. 2019. Doubly robust conditional logistic regression. *Stat. Med.* 38(23):4749–60