Simpson's Paradox Beyond Confounding

Zili Dong[1], Weixin Cai[2], Shimin Zhao[3]

Forthcoming in *European Journal for Philosophy of Science*

Abstract: Simpson's paradox (SP) is a statistical phenomenon where the association between two variables reverses, disappears, or emerges, after conditioning on a third variable. It has been proposed (by, e.g., Judea Pearl) that SP should be analyzed using the framework of graphical causal models (i.e., causal DAGs) in which SP is diagnosed as a symptom of confounding bias. This paper contends that this confounding-based analysis cannot fully capture SP: there are cases of SP that cannot be explained away in terms of confounding. Previous works have argued that some cases of SP do not require causal analysis at all. Despite being a logically valid counterexample, we argue that this type of cases poses only a limited challenge to Pearl's analysis of SP. In our view, a more powerful challenge to Pearl comes from cases of SP that do require causal analysis but can arise without confounding. We demonstrate with examples that accidental associations due to genetic drift, the use of inappropriate aggregate variables as causes, and interactions between units (i.e., inter-unit causation) can all give rise to SP of this type. The discussion is also extended to the amalgamation paradox (of which SP is a special form) which can occur due to the use of non-collapsible association measures, in the absence of confounding.

[1] Independent scholar.
[2] Department of Philosophy, University of California San Diego, 9500 Gilman Dr., LaJolla, CA 92093, USA.
[3] Department of Philosophy, University of Wisconsin-Madison, 1300 University Ave., Madison, WI 53707, USA.

Keywords: Simpson's paradox, causal modelling, DAGs, confounding[4]

1. Introduction

Simpson's paradox (SP), in its most striking form, is a phenomenon where a statistical association between two variables *X* and *Y* in the entire group (or aggregate data) reverses in every sub-group (or disaggregate data). SP also includes cases where the association between *X* and *Y* in the entire group disappears in each sub-group, as well as cases where an association between *X* and *Y* appears in each sub-group even if they are unassociated in the whole group. Since we typically partition a group based on some third variable of interest *Z* (e.g., *Sex*), we may also say that SP occurs when the association between *X* and *Y reverses*, *disappears*, or *emerges*, after conditioning on *Z*. For a recent survey of SP, see Sprenger and Weinberger (2021).

Two clarifications on the above definition of SP are needed. Firstly, following Hoover (2003) and Sprenger and Weinberger (2021), we understand SP primarily as a sample-level phenomenon that can be readily observed in statistical data. In this paper, the concept of sample *association* is distinguished from the concept of population (or probabilistic) *correlation*. Of course, if we have a case of SP in which sample associations between *X*, *Y*, and *Z* adequately indicate the probabilistic correlations between them, this will be a case of SP defined in terms of both association and correlation.[5] Secondly, even though we agree that causality plays a key role

[4] We thank Xiuyuan An, Holly Andersen, Nancy Cartwright, David Danks, Yichen Luo, Tianqin Ren, Elliott Sober, Michael Titelbaum, Anqi Wang, and two anonymous reviewers for their helpful suggestions on earlier versions of this paper.

[5] In most cases, we can safely ignore the difference between sample associations and probabilistic correlations. Still, there are cases in which it is important that we separate them, as we shall see later in the paper.

in understanding many important cases of SP, the definition of SP we adopt in this paper does *not* stipulate that SP must be a causal phenomenon.[6]

Probably the most well-known example of SP is the case of graduate admissions at the University of California, Berkeley in 1973 (Bickel, Hammel & O'Connell, 1975). It was recorded that at the university level, about 44 percent of the males and about 35 percent of the females were admitted. This means that being female was negatively associated with being admitted to the University of California, Berkeley, which suggests that there might have been discrimination against female applicants in the admissions process. However, if we break down the data, we will find that, in the majority of the departments, there was no significant bias against female applicants. In fact, in a few departments, females were even more likely to be admitted than males. This poses the question of whether there was truly sex discrimination that affected the admissions committee's decisions. That is, if we want to identify the existence of sex bias in the admissions process, should we look at the university-level data or the department-level data?

Although SP may seem like a purely statistical or probabilistic oddity, philosophers have long recognized its causal roots. Back in the 1980s, SP was posed as an important challenge to probability-raising accounts of causality (see, e.g., Cartwright, 1979). In this context, Cartwright rightly pointed out that the association between the cause-variable X and the third variable Z is essential for the occurrence of SP.[7] To avoid SP, she proposed that we measure causal effects relative to the so-called 'causally homogeneous' populations or reference classes in which there is

[6] It has been suggested that SP, in its nature, is a *causal* phenomenon (e.g., Pearl, 2014; we shall come back to this later in the paper). For Pearl, a genuine case of SP must be embedded in a causal context. Apparent cases of SP that lack causal context are dismissed by him as not genuinely paradoxical (he calls such cases "Simpson's reversal"). Although we think the distinction Pearl draws between Simpson's paradox and Simpson's reversal is well-motivated, we also find it somewhat *ad hoc* to stipulate that SP must be a causal phenomenon. We show in Section 3 that there are more principled reasons why cases of SP that lack causal context should be distinguished from those cases that have a causal context, without having to draw the distinction by stipulation.

[7] See also Sprenger and Weinberger (2021) for a detailed explanation of why an association between X and the third variable Z is *necessary* for SP to occur.

no association between *X* and *Z*, since *Z* can be seen to have been 'held fixed' in such homogeneous reference classes. However, this proposal, embedded in the probability-raising approach to causality, is subject to the approach's inability to explicitly represent causal structures underlying the SP. As early as 1987, Irzik and Meyer had realized the inadequacy of Cartwright's solution and suggested we analyze the causal structure of SP using tools of causal modelling (Irzik & Meyer, 1987).[8] The tool they used is the method of path analysis (invented by Sewall Wright around 1920), which is a precursor to the more powerful framework of graphical causal modelling developed later by Spirtes et al. (2000) and Pearl (2009).

Proponents of the framework of graphical causal models propose to analyze SP in *causal-graphical* terms (Pearl, 2009, 2014; Pearl et al., 2016; Pearl & Mackenzie, 2018; Spirtes et al., 2000). Notably, Pearl et al. (2016) assert that they can "fully resolve Simpson's Paradox by determining which variables to measure and how to estimate causal effects under confounding" (p. 44). For Pearl et al., *confounding* is to be analyzed in terms of causally interpreted directed acyclic graphs (DAGs). Essentially, their proposal is that for *all* types of SP involving *X*, *Y*, and *Z*, the paradox can be resolved by construing the partitioning variable *Z* as a confounding variable relative to a causal DAG over {*X*, *Y*, *Z*}. We will explain this in more details in Section 2 (see especially Figure 1 there).

In this paper, we contend that this causal-graphical analysis of SP, despite offering genuine insight, is not complete and cannot fully resolve SP as Pearl et al. have claimed. We acknowledge that many important types of SP do arise from confounding; however, we do not think this is a universal feature of SP. For one thing, some cases of SP lack causal context, as has been argued by Bandyopadhyay et al. (2015). While acknowledging that Bandyopadhyay et al. raise a logically

[8] Irzik and Meyer's (1987) analysis of SP is remarkably farsighted; its core idea is basically the same as Pearl's confounding-based analysis of SP (see the illustrative example they give on p. 513).

valid objection against the completeness of Pearl's causal-graphical analysis of SP, we will examine their argument from the perspective of scientific and statistical practice and show that, at least from this perspective, it poses no significant challenge to Pearl's analysis. More importantly, we show that even if we narrow down our attention to cases of SP that do need a causal treatment, we can still find various types of SP that do *not* involve confounding of any sort. After carefully examining these types of SP, we conclude that SP should be seen as a symptom with many aetiologies and there appears to be no unified analysis that can capture all of them.

The rest of the article proceeds as follows. In Section 2, we first present Pearl's causal-graphical analysis of SP and then point out its limitations which are to be further examined in subsequent sections. In Section 3, we argue that Bandyopadhyay et al.'s alleged counterexample to Pearl's analysis (i.e., the Marble example) does not constitute a threatening objection to Pearl, because the example presupposes strong accidental associations which are unlikely to be encountered in ordinary life or scientific practices. Section 4 discusses three cases of SP which all require a causal analysis, but Pearl's analysis does not apply. Specifically, in Section 4.1, we discuss a case of SP arising in an evolutionary context due to accidental associations (random genetic drift). Section 4.2 discusses a case of SP arising from the use of inappropriate aggregate variables as causes, and Section 4.3 discusses a case of SP arising from inter-unit causation (illustrated using a non-stationary time series example). Section 5 extends the discussion to a generalized version of 'SP-type' phenomena known as the amalgamation paradox (AP).[9] It has been recognized by epidemiologists—but less known by philosophers—that some cases of AP can

[9] It is sometimes said that AP is "the most generalized version of [SP]" (Sprenger & Weinberger, 2021), and sometimes SP and AP are used as synonyms (as in, e.g., Hernán, Clayton & Keiding, 2011). Admittedly, this use of terminology is confusing. To avoid confusion, we make a clear distinction between SP and AP in this paper: AP is defined as a broader category than SP, with SP being a special case of AP. This distinction will be important for our discussion in Section 5.

occur without confounding; for this reason, we note that Pearl's analysis of SP cannot be generalized to AP. Section 6 is a brief conclusion.

2. The scope and limitations of the causal-graphical analysis

In this article, we focus on Pearl's (especially, 2014) causal-graphical analysis of SP since, to our knowledge, this is the most influential and systematic treatment of SP by far. At the centre of his analysis is the framework of graphical causal models. The first thing to note is that although Pearl's analysis of SP is often referred to as the 'causal analysis', it is more accurate to call it the 'causal-*graphical* analysis'. This is not merely a verbal issue: while graphical modelling is undoubtedly an important tool in causal inference, a graphical model by no means captures everything interesting about a system's causal properties (Cartwright, 2001; Dawid, 2010).

Directed acyclic graphs (DAGs) are the most often used type of graphical models in causal inference. A causally interpreted DAG G consists of a set of vertices, which represent a set of causal variables $\mathbf{V} = \{X, Y, Z, \ldots\}$, and a set of edges, which represent *direct* causal relations between the variables. X is a direct cause of Y (relative to G) in the sense that it is possible to change the value of X through some atomic or ideal interventions such that the probability distribution of Y will change accordingly, when all other variables in the graph are held fixed by interventions (Pearl, 2009; Woodward, 2003). G is directed, which means that all the edges are single-headed arrows. We define a causal path between X and Y on G as a path on which the edges between X and Y are all directed in the same direction (e.g., $X \rightarrow Z \rightarrow W \rightarrow Y$). G is also acyclic, which means that it contains no causal circle (i.e., causal path that starts and ends with the same variable, e.g., $X \rightarrow Z \rightarrow X$). Besides, G and its corresponding joint probabilistic distribution over $\mathbf{V}$ are assumed to satisfy the *causal Markov condition* (a modern successor of Reichenbach's principle of common cause). This condition says that for any variable X in $\mathbf{V}$, X is probabilistically

independent of every other variable except *X*'s descendants (i.e., *X*'s effects), conditional on *X*'s parents (i.e., *X*'s direct causes).

DAGs have proven empirically fruitful in representing and analyzing the causal structure or data-generating process relevant to a causal investigation. They are particularly suited for handling the notorious problem of *confounding* in causal inference (Pearl et al., 2016; Shrier & Platt, 2008).[10] The primary task of causal inference in many scientific domains (especially in high-level sciences such as biology, sociology, epidemiology, etc.) is to identify and estimate causal effects. Confounding is an important type of systematic source of error that might occur during causal effect estimation. An error is systematic means that its occurrence is not accidental; that is, it does not occur by chance. For example, if we use the marginal association between *X* and *Y* in the observational data to measure the direct effect of *X* on *Y*, we may misestimate the effect when there are indirect causal paths between *X* and *Y* that bring about indirect effects of *X* on *Y*. For Pearl, such kind of discrepancy should be assessed using DAGs which can visually represent all the relevant causal assumptions. With these causal assumptions and the help of certain graphical rules (e.g., the back-door criterion), we can then eliminate confounding by adjusting for or conditioning on a (sufficient) set of confounding variables (Pearl et al., 2016).

For our purposes below, it suffices to focus on the estimation of the direct effect of *X* on *Y* relative to a pre-specified DAG on {*X*, *Y*, *Z*}. Relative to a DAG on {*X*, *Y*, *Z*}, our estimation of the direct effect of *X* on *Y* will be confounded if and only if any of the following situations obtains: (a) we fail to adjust for *Z* when *Z* is a common cause of *X* and *Y*, (b) we mistakenly condition on *Z* when it is a common effect of *X* and *Y* (here *Z* is also called a 'collider' and this type of

[10] For various reasons, the term 'confounding' has been used in confounding ways (pun intended). For example, sometimes 'confounding' is used to refer merely to bias due to lurking common causes (cf. Hernan et al. 2011). Our usage here is much broader, following Pearl, Glymour, and Jewell (2016). On this broad usage, a confounding variable may also be a collider or a mediator; see our discussion below.

confounding is also known as 'collider bias'), or (c) we fail to adjust for *Z* when *Z* is a mediator between *X* and *Y* (i.e., *Z* is on an indirect causal path between *X* and *Y*). Figure 1 illustrates these three cases of confounding with DAGs.

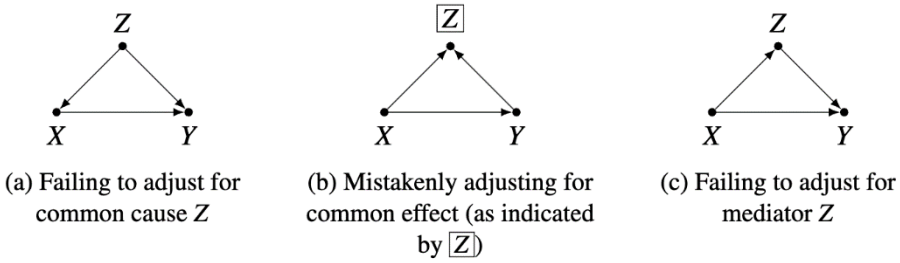


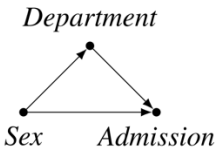
Figure 1 Three possible cases of confounding in the estimation of the direct effect of *X* on *Y*.

Consider a concrete example. If we are interested in the effect of Paxlovid pills in reducing deaths from COVID-19, we cannot simply compare the death rate in a group of COVID-19 patients who took the pills with the death rate in another group who did not. This estimation of the drug's effectiveness is confounded since patients at a higher risk of dying from COVID-19 are also much more likely to receive the pills. In other words, risk factors for Covid-19 such as age or chronic diseases, if unadjusted for, will lead to confounding (of type (a) mentioned above) in the effect estimation.

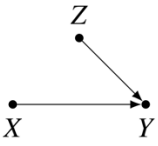
Pearl (2014) suggests that we use DAGs to analyze causal structures that are at work behind various cases of SP. In his analysis, SP is diagnosed as a peculiar consequence of confounding. The basic idea is the following. The reason we find SP 'paradoxical' is that we think how *X* (e.g., *Sex*) affects *Y* (e.g., *Admission*) should not depend on the level at which the influence is measured. If the associations between *X* and *Y* in the large group and in the sub-groups disagree with each other, they cannot both indicate the 'true' effect of *X* on *Y*—at least one of the associations must be 'spurious'. For it violates our causal intuition to say that depending on the way we look at the

data, the applicant's sex can both influence, and not influence, admissions. Pearl's key insight is that the kind of spurious association that leads to SP should be seen as resulting from confounding. If we can eliminate the confounding responsible for a case of SP (assuming that the relevant causal structure is known), we will obtain an unconfounded estimation of the true effect of interest. The paradoxicality and counter-intuitiveness of SP will then be explained away.

Pearl (2014) identifies a group of causal structures that can give rise to SP, together with a group of causal structures that cannot. Here we consider one example representative of each type. First, consider the DAG in Figure 2a, which depicts the putative causal structure responsible for the example of Berkeley's graduate admissions.[11] Note that there are two causal paths from *Sex* to *Admission*, which means that, relative to this DAG, *Sex* is both a direct and an indirect cause of *Admission*. This type of causal structure is prone to generating SP, because the causal influence that *Sex* has on *Admission* along the direct causal path and the indirect one could be in opposite directions.



(a) The causal structure for Berkeley's graduate admission example



(b) A causal structure that (allegedly) cannot generate SP, according to Pearl

Figure 2 A comparison of a causal structure that can generate SP and a causal structure that (allegedly) cannot generate SP.

[11] This causal graph is adapted from Pearl and Mackenzie (2018, p. 312, Fig. 9.4). Pearl and Mackenzie also considered more complicated causal structures, but these complications are not essential for our discussion here.

More specifically, *Sex* will have a direct effect on *Admission* if sex bias does exist in the admissions process (e.g., being female causes one to be discriminated against in this process).[12] At the same time, one's sex may influence one's department choice, which can further affect one's chance of being admitted by the university because some departments are harder to get in than others. Thus, *Sex* also has an indirect effect on *Admission*, which is mediated by *Department*. In this case, the association between *Sex* and *Admission* in the university-level data, without adjusting for *Department*, gives us an estimation of the sum of the direct and indirect effect (i.e., the *total* effect) of *Sex* on *Admission*. To examine whether there is sex bias against female applicants—that is, whether *Sex* has a *direct* effect on *Admission*—we should instead look at the association between *Sex* and *Admission* when *Department* is adjusted for, given that the decision to admit an applicant was processed within each department. In other words, we should rely on the association between *Sex* and *Admission* in the department-level data to infer whether female applicants were discriminated against in the admissions process. If one tries to identify sex discrimination using university-level data, the estimation will be confounded.

In contrast, the causal structure in Figure 2b is claimed by Pearl to be unable to give rise to SP. His reason is that, in this DAG, *Z* cannot bring about any of the three aforementioned types of confounding: *Z* is not a common cause, a collider, or a mediator. Given the causal Markov condition, the DAG implies that *X* and *Z* are *probabilistically* independent. Under the assumption that probabilistic independence implies statistical independence, *X* and *Z* should be found unassociated in the data. It is this assumed absence of association between *X* and *Z* that justifies Pearl's claim that SP will not arise for this DAG. However, this assumption does not always hold. It is still possible that *X* and *Z*, despite being causally and probabilistically independent, are

[12] Note that the direct effect of *Sex* on *Admission* only reflects possible sex biases during the admissions process. This cannot tell us anything about structural sexism.

accidentally associated in the data we collected, especially when the sample size is small; in such cases, we may still encounter SP in the absence of confounding. Fortunately, this type of situation is unlikely to arise in sufficiently large samples. Additionally, note that both associations arising from chance and associations resulting from confounding are 'spurious' in the sense that they do not reflect true causal effects. The difference between them, however, is that in statistical practices, the latter is much more robust and common. This is also why Pearl simply disregards the possibility of accidental associations in his discussion. We will come back to these points in Section 3.

Undoubtedly, Pearl's causal-graphical analysis of SP contains genuine insight and covers a broad range of cases of SP. That said, we believe this analysis is inadequate in important ways. His analysis focuses on those cases of SP in which the associations involving X, Y, and Z are supposed to be explained by positing an underlying causal structure over $\{X, Y, Z\}$, represented using a DAG. However, it is well-known that DAGs make certain assumptions about the causal structures they represent which do not necessarily hold in practice.[13] In particular, a standard DAG presupposes at least two things. Firstly, the causal variables have been well-chosen. Poor choices of variables, for instance, using variables that are inappropriately defined or logically connected, may bring trouble to causal inference (see Woodward, 2016). Secondly, an association between two variables can always be causally explained, either as a result of genuine causation or as that of confounding.[14] As we shall see later, these assumptions, while making DAGs expedient to use, may not always hold.

[13] Broadly speaking, our point here aligns with Spanos' claim that Pearl's "causal explanation of the paradox largely ignores some of these empirical issues by viewing it as a purely probabilistic conundrum … when models are estimated using actual data, one needs to secure the validity of the model assumptions vis-a-vis the data before any causal information can be utilized reliably" (Spanos, 2021, p. 608). In other words, investigators run the risk of introducing unwarranted causal information into the model when those model assumptions are not examined for their validity.

[14] It must also be acknowledged that these assumptions may be made by other approaches to causal inference as well. For example, that causal variables should be in some sense well-chosen is a prerequisite not just for graphical causal modelling but also for the potential-outcomes approach to causal inference.

It follows that Pearl's causal-graphical analysis of SP will not apply in either of the following two types of cases:

> **Case 1:** The occurrence of SP does not require the existence of any causal relationships among the relevant variables (i.e., *X*, *Y*, and *Z*). That is, none of the associations between *X*, *Y*, and *Z* in this case of SP needs to be generated by causation among these variables. To make sense of this type of SP, causal information is irrelevant.

> **Case 2:** SP occurs in a context in which the effect of *X* on *Y* is queried, prompting an investigation into the causal story underlying SP. Moreover, this causal story is required for explaining away SP in this case. However, the presence of this type of SP does not result from confounding; its root lies somewhere else.

The sort of counterexample to Pearl's analysis in **Case 1** has been extensively examined by Bandyopadhyay et al. (2015), which will be the focus of our discussion in the next section. As Bandyopadhyay et al. argue, at least in some cases, SP need not be analyzed in causal terms. We agree with them that their counterexample shows that Pearl's analysis cannot adequately account for *all* possible instances of SP. Nevertheless, we believe it is equally important to explore counterexamples in the line of **Case 2**. For one thing, while Bandyopadhyay et al.'s counterexample, as an instance of **Case 1**, constitutes a *logically* valid objection to Pearl's analysis, we believe that it is not a *practically* strong challenge. As we shall explain, the occurrence of their counterexample is extremely rare in ordinary and scientific contexts. In our view, it is

counterexamples in the line of **Case 2** that offer a stronger ground against the adequacy of Pearl's analysis.

For another, **Case 2** has received limited attention in the literature, although the question of whether there are causal grounds of SP that do not involve confounding and thus are not subject to Pearl's analysis is evidently important. Pearl's causal-graphical analysis regards confounding as an indispensable condition for SP and considers confounding adjustment as sufficient for resolving SP. This, if true, implies that practitioners of causal inference need not worry about possible instances of SP if no confounding is present. In this paper, we show that this is not the case. SP can arise in causal contexts where confounding is absent. In Section 4, we offer three cases of SP that fall under **Case 2**. As we shall see, for all three cases, although causal information plays a key role in explaining away the paradox, none of them involves confounding.

3. Simpson's paradox without causation? A statistical-practice perspective

Bandyopadhyay et al. (2011, 2015) (and more recently, Sarkar & Bandyopadhyay, 2021) defend a non-causal, "*logic-based*" analysis of SP. According to them, at least in some cases, SP is fundamentally an arithmetic oddity, whose nature has nothing to do with causality. In their view, SP "involves the reversal of the direction of a comparison or the cessation of an association when data from several sets are pooled" (Bandyopadhyay et al., 2015, p. 13; note that this is essentially equivalent to the definition of SP we give at the beginning of this paper). Importantly, Bandyopadhyay et al. emphasize that the satisfaction of these conditions need not presuppose any causal relations among X, Y, and Z. Moreover, causal information is also unnecessary for explaining why data patterns satisfying these conditions can give us a feeling of puzzlement (and thus be regarded as "paradoxical").

Therefore, Bandyopadhyay et al. claim, it is not the case that SP can only be explained away by positing causal relations among the relevant variables. Note that they do not deny that sometimes we need to analyze SP causally. However, "SP has to do with causality *only if* we ask the *what-to-do* question"—the question of "What should one *do* when confronted with a typical case of the paradox?" (Bandyopadhyay et al., 2015, p. 14; emphasis added). In other words, the causal structure underlying an instance of SP becomes relevant only in circumstances where one is considering what decision should be made in view of the associations exhibited in this instance of SP. For example, when we encounter a situation where it appears that the effect of a treatment reverses after conditioning on sex, the question of how such 'contradictory' evidence should guide the use of the treatment arises. In cases where no such decision-making concern arises, according to Bandyopadhyay et al., there would be no need to understand SP through a causal lens.

Red Marble Rates	**Bag 1**	**Bag 2**	**Total**
Big Marbles	180/200 = **90%**	100/300 = **33%**	280/500 = 56%
Small Marbles	480/600 = 80%	10/100 = 10%	490/700 = **70%**

Table 1 The Marble example with fictitious data (adapted from Bandyopadhyay et al., 2015, Table 5).

Bandyopadhyay et al. use the Marble example to illustrate their point. Suppose there are two bags of marbles with different sizes and colours. Table 1 reports how many marbles of a specific size-colour pair there are in each bag. It is observed that within both bags (Bag 1 and Bag 2), big marbles are more likely to be red, compared with small marbles. Given this, it seems reasonable to expect that the same pattern will hold once we merge all the marbles together. However, this expectation cannot be fulfilled: the association between *Size* and *Colour* reverses in the aggregate data. We find this result surprising, peculiar, or perplexing because it violates our

expectation. According to Bandyopadhyay et al., "[t]here are no causal assumptions made in this example" regarding how *Size*, *Colour*, and *Bag* are related to one another (Bandyopadhyay et al., 2015, p. 19). This implies that no confounding can be appealed to in accounting for why this instance of SP occurs. Granted that this is the case, Bandyopadhyay et al. contend that at least in this example, SP has a purely arithmetic root, which implies that "SP is not basically causal" (p. 13). Thus, Pearl's causal-graphical analysis cannot be a universal treatment of SP.

We agree that Bandyopadhyay et al.'s Marble example is a genuine case of SP: the example does involve the kind of association reversal characteristic of SP, and therefore, conforms to the definition of SP we give at the beginning of this paper.[15] In addition, we agree that the example does stimulate our intuitive perplexity and that the perplexity can be explained away without invoking any causal postulates involving the relevant variables. Indeed, the example is designed in this way in order to constitute a counterexample to Pearl's causal-graphical analysis.

Nevertheless, we find the Marble example inadequate in an important sense. From a *statistical-practice* perspective (as opposed to a purely 'logic-based' or 'arithmetic' one), this example presupposes a highly uneven distribution of marbles in the two bags. In particular, in Bag 1, ¾ of the marbles are small, whereas in Bag 2, ¾ of the marbles are big, which means there is a *strong association* between *Size* and *Bag*. This distribution of the marbles is extremely rare from a statistical practice perspective. This is because, assuming that the causal Markov condition is satisfied, these variables are not expected to be found highly associated in almost all the data we

[15] Pearl would give a different response to the Marble example: he would think that this is not a genuine case of SP at all but merely a case of Simpson's 'Reversal' (see footnote 6). For Pearl, it is essential for a genuine case of SP that it violates our *causal* intuition. If we understand him correctly, he seems to think that the violation of the causal intuition (as we talked about in Section 2) should be a *defining* feature of SP. Although we agree with Pearl that cases of SP that violate this causal intuition are the exemplars of SP, and we also agree on the explanatory significance of this causal intuition, we hesitate to make this feature definitional of SP. After all, the definition of SP we give in Section 1—which is also widely used in the literature—is not formulated in causal terms. Besides, many of our concepts have both typical and atypical cases. The Marble example seems to be better conceived as a less typical case of SP.

may encounter in practice (given the relatively large sample size in this example) unless we posit causal connections among them. In fact, under the hypothesis that there is no causation (and thus no correlation) between *Size* and *Bag* in the Marble example, the probability of observing a difference in the proportions of big marbles between the two bags is 50% (i.e., 300/400 – 200/800) or more is lower than 0.000001: it happens less than once in a million times. Similarly, the strong association between *Colour* and *Bag* presented in Table 1 is also extremely rare, assuming that *Colour* and *Bag* are not correlated.

Note that Bandyopadhyay et al. cannot explain the strong association between *Size* and *Bag* and that between *Colour* and *Bag* in Table 1 by appealing to a biased process of disproportionally sorting more small marbles and red marbles into Bag 1. This is because, in that case, the Marble example would require a *causal* analysis, since this biased sampling process is actually an instance of collider bias (see Section 2 for more on collider bias). That is, if the size and colour of a marble affect which bag it will be put into, then *Bag* will become a collider whose value is affected by both *Size* and *Colour*. Now that a causal structure behind the data has been introduced, the Marble example will no longer be a counterexample to Pearl's causal-graphical analysis.

Therefore, the associations in the Marble example can only be 'chancy' or *accidental* in the sense that they are a result of very rare random fluctuations in the process of sorting marbles into bags. These accidental associations, especially the one between *Size* and *Bag*, play a key role in generating SP in this example: if we were to render the proportions of big marbles in both bags roughly the same, SP would not occur. The fact that the occurrence of SP in the Marble example depends on strong accidental associations implies that, *statistically* speaking, this kind of SP (**Case 1**) is far less common or robust, compared to typical cases of SP such as the Berkeley admissions

example. We have very little reason to expect that data patterns similar to the Marble example will be ever observed in statistical practice. By contrast, in the Berkeley example, there is a causal structure that *robustly* generates associations among *Sex*, *Department*, and *Admission*. This causal structure may also be found in other years or places. For this reason, we may expect to find similar occurrences of SP in the admissions processes in the years after 1973 at UC Berkeley or other universities.

We think Bandyopadhyay et al.'s discussion of SP does not give due emphasis to the fact that instances of SP generated by robust causal structures and those generated by chancy fluctuations are *not* on a par with respect to their importance in scientific practice. In most contexts, scientists collect and analyze data in order to reveal underlying causal structures (or for other causally relevant goals, including explanation, prediction, understanding, and control).[16] The connection between a causal structure and the associations it robustly generates allows scientists to infer the former from the latter. However, when associations seem to provide 'contradictory' evidence for the underlying causal structure, as we saw in the Berkeley admissions example, it can be particularly puzzling; this is why SP as a statistical phenomenon has received so much scientific attention.

Importantly, since accidental associations may mislead causal inference, scientists will take various measures (e.g., collecting more data, and conducting significance tests) in attempts to reduce the possibility that the evidence they observe comes from an accidental association in a particular dataset. These measures enable scientists to quickly discover and discard, to their best

[16] Notably, there are data-generating processes that are non-causal but remain scientifically significant. For example, nonlocal quantum correlations that violate Bell inequalities are usually considered not subject to a causal interpretation (cf. Myrvold et al., 2024). Despite this, these correlations have been robustly observed in well-designed Bell experiments, which is why they are of great scientific significance. See Frisch (2020) for a brief survey of this issue; see Wood & Spekkens (2015) and Näger (2022) for recent discussions. However, because the literature on SP has almost exclusively focused on high-level sciences where quantum effects can be safely ignored, we will also set aside quantum correlations in this paper. We thank an anonymous reviewer for suggesting this point.

knowledge, accidental associations due to chancy fluctuations in sampling. For this reason, SP arising from accidental associations is of minimal scientific relevance.[17] Therefore, we claim, it is mainly through a causal lens that scientists come to see cases of SP as intellectually intriguing and practically significant. In contrast, although the Marble example successfully demonstrates that Pearl's causal-graphical analysis is not a universally valid treatment of SP, its dependence on an extremely rare random fluctuation precludes it from being a scientifically and practically significant counterexample.

Lastly, we disagree with Bandyopadhyay et al. (2015) that causality only matters when decision-making questions are asked regarding a case of SP. One may still want to conduct a causal inquiry about how an instance of SP is generated when one's goal is purely *epistemic*, such as explanation or understanding. It is quite often that people scrutinize data merely for the sake of gaining an understanding of its underlying data-generating process, without immediate concern for decision-making. For instance, one might have a purely epistemic interest in the historical question of whether there was truly sex bias in Berkeley's 1973 graduate admissions process, even if she does not want to make any decisions.

Interestingly, we do find a type of SP that arises due to an accidental association between *X* and *Z*, and at the same time, has a more visible scientific significance. Unlike the Marble example, however, this new type of SP needs to postulate causal relations among the relevant variables. So, these two types of SP, despite both relying on accidental associations, are importantly different:

[17] Note that we are not denying the value of investigating 'outliers' or 'exceptional' data points. As pointed out by an anonymous reviewer, there are ample examples in medicine where paying attention to 'exceptional responses' to a treatment in clinical trials led to important medical progress (Mukherjee, 2015). However, as Mukherjee also made clear, the rarity of these cases is due to complex interactions between numerous *causal* factors. These factors are of scientific interest because if we can discover and control these factors, we will be able to robustly generate these 'exceptional' cases. Therefore, they are not 'chancy' fluctuations in the statistical sense which are not generated by any interesting causal factors.

one is situated in a causal context whereas the other need not be. This is why we will leave the discussion to Section 4.

4. Simpson's paradox with causal contexts but beyond confounding

This section aims to demonstrate that even among cases of SP that are embedded in a causal context, some of them cannot be analyzed as a consequence of confounding. Note that we are not saying that these cases should not be analyzed causally, nor are we saying DAGs are no longer useful for analyzing them. Our claim is specifically that these cases need to be treated in a careful and nuanced manner which goes beyond what Pearl's (2014) analysis can offer.

4.1. Accidental associations

The previous section discussed a form of SP free of causal context, as illustrated using the Marble example. In this section, we shall see that similar to the Marble example, accidental associations can also play a key role in generating SP when the instance of SP is situated in a causal context. More specifically, an accidental association between X and Z can generate SP over $\{X, Y, Z\}$ when they form a causal structure like the following: $X \rightarrow Y \leftarrow Z$ (see Figure 2b). Recall that in Section 2, we noted that this causal structure cannot generate SP, *only* on the assumption that there is no accidental association between X and Z. But if X and Z are, in fact, accidentally associated, SP can still arise in this causal structure.

Consider the following example (see Table 2; adapted from Sober, 2024, p. 44) about how being selfish or altruistic may affect the expected number of offspring of an individual (i.e., fitness). There are two groups, A and B, each containing two traits: being selfish, or being altruistic.[18] A trait's fitness in a group is defined by how many offspring individuals of the trait in that group produce on average. For example, in group A, there are 200 selfish individuals, and they have 800

[18] For the sake of simplicity, let us assume asexual reproduction, no mutation, and no migration.

offspring in total, so the fitness of selfishness in group A is 800/200=4. As shown in the table, being selfish is positively associated with fitness in both group A and group B. However, the association between selfishness and fitness reverses once the two groups are analyzed as a whole (i.e., as a metapopulation). This is paradoxical if we interpret fitness as representing the *causal* propensity of a trait in producing offspring.

Fitnesses	**Group A**	**Group B**	**Total**
Selfish	800/200 = **4**	1600/800 = **2**	2400/1000 = 2.4
Altruistic	2400/800 = 3	200/200 = 1	2600/1000 =**2.6**

Table 2 A hypothetical example of SP involving the fitness of being selfish or altruistic.

The paradoxicality of this example is primarily due to the existence of a very strong association between variables *Selfish* and *Group*: in our example, group A is dominantly altruistic whereas group B is dominantly selfish. Without this association, SP would not be possible. But where does this association come from? It might be because having a particular trait *causes* one to be in a certain group—making this case of SP similar to the Berkeley example.[19] But what we want to show below is that the association need not necessarily come from a causal source.

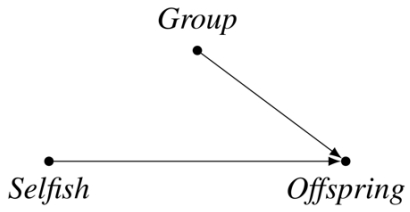


Figure 3 A causal graph for the SP involving *Selfish*, *Group* and *Offspring*, without positing a causal relation between *Selfish* and *Group*.

[19] This situation also seems to be what Sober (2024, Sect. 3.2) has in mind.

Even if we assume that *Selfish* and *Group* are causally *in*dependent (as depicted in Figure 3**Figure *3***), we may still be able to accidentally observe a strong association between them. Regarding our example specifically, such a strong association may have resulted from genetic drift, such as the founder effect (Dobzhansky & Pavlovsky, 1957). That is, by pure chance, two relatively small groups of individuals may happen to have highly uneven distributions of altruism and selfishness upon being separated from the larger group. Put differently, the association between *Selfish* and *Group* in the above example can be seen as a result of random fluctuations.

Clearly, this chancy association is not a result of confounding, since *Selfish* and *Group* have been assumed to be causally independent. Still, this accidental association can mislead causal effect estimation and lead to SP. Therefore, if we are to estimate the fitness of selfishness reliably, we need to first make sure that genetic drift does not generate a significant accidental association between *Selfish* and *Group* in the data.

4.2. Aggregate variables

An important type of scenario that can supply 'paradoxical' associations needed for SP but cannot be analyzed away in terms of confounding involves the use of aggregate or summed variables as causes. An aggregate variable is one that can be written as the sum of two or more other variables (e.g., $X=X_1+X_2$). It is known that an aggregate variable may have an *ambiguous* effect on an outcome of interest, if the variables it sums up have heterogenous effects on the outcome (Spirtes & Scheines, 2004). A well-known example of such kind of aggregate variable is total cholesterol (*TC*). *TC* consists of both low-density lipoprotein (*LDL*) and high-density lipoprotein (*HDL*), which have opposite effects on cardiovascular diseases (*CVD*). The ambiguity in the effects of such kind of aggregate variables, as we shall see, opens a door for SP.

Imagine that we have a group of patients in which levels of total cholesterol are *unassociated* with the incidence of cardiovascular diseases, which suggests the prima facie result that *TC* has no effect on *CVD*. However, if we condition on levels of *HDL*, surprisingly, *TC* becomes *positively* associated with *CVD* (i.e., a higher level of *TC* is associated with a higher *CVD* rate, conditional on *HDL*). So, a positive association emerges upon conditioning the third variable—making this a case of SP. Table 3 represents the results of a fictitious dataset for this example. It shows that *CVD* is unassociated with levels of *TC* in the entire study group (50% vs. 50%). However, within both sub-groups, having a high level of *TC* seems to make the patients more likely to develop cardiovascular disease. How should we explain this paradoxical result?

***CVD* Rates**	**Low *HDL***	**High *HDL***	**Total**
Low *TC*	40/70 = 57.1%	20/50 = 40%	60/120 = 50%
High *TC*	80/130 = **61.5%**	120/270 = **44.4%**	200/400 = 50%

Table 3 Fictitious data for the incidence of cardiovascular disease among sub-groups (with low and high levels of *HDL*) and among the whole group.

Due to the *non-causal* (specifically, logical) relationship among *TC*, *HDL*, and *LDL* (i.e., *TC=LDL+HDL*), including these three variables in a single *causal* graph may cause trouble. So, to say the least, caution is needed if one attempts to provide a graphical analysis of the above example. In particular, it is helpful to represent non-causal relationships using dashed arrows so as to distinguish them from causal relationships (see Figure 4). Besides, in the DAG we draw, there is no need to draw a causal arrow from *TC* to *CVD*, given that all the work *TC* appears to be doing is in fact done by *LDL* and *HDL*.

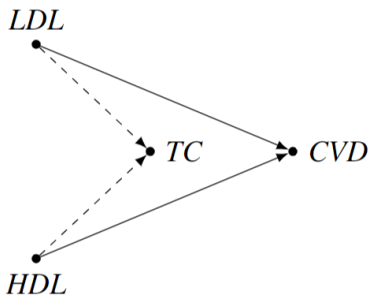


Figure 4 A 'mixed' DAG containing both non-causal relations and causal relations.

However, the sort of mixed graph in Figure 4 is nothing like a standard DAG. It is widely acknowledged and observed in causal inference practice that variables in a standard DAG should not stand in non-causal relationships. Therefore, there is no standard DAG representation for the causal structure underlying the type of SP we are discussing here. In particular, *HDL* in Figure 4 is *not* a confounding variable since it is not even a cause of *TC*. This means that Pearl's causal-graphical analysis cannot handle this type of SP.

How should we explain away and avoid this type of paradox then? The answer is simple: *TC* should not be used as a cause of *CVD* to begin with, because its 'effect' on *CVD* is ambiguous. Instead, we should use *HDL* and *LDL* as causes when investigating the incidence of cardiovascular diseases. So, it turns out that the genuine source of SP in this case is not confounding but a bad choice of causal variables, that is, the use of inappropriately summed variables as a cause.

4.3. Inter-unit causation

Following the broader point behind Section 4.2, we now present another type of SP that can arise due to inappropriate variable choice instead of confounding. This type of SP has its root in having chosen a set of variables which are defined on a group of units that causally interact with each other. This phenomenon is known as *inter-unit causation* (Spirtes et al., 2000, p. 296; J. Zhang & Spirtes, 2014) or interactions among units (C. Zhang et al., 2022). As we shall see, inter-

unit causation may create an association between two causally independent variables, giving rise to SP. This type of SP is not due to confounding, and thus falls outside of Pearl's (2014) analysis; a more nuanced causal analysis is needed.

Such type of SP can be found, for example, in non-stationary time series data, a series of data whose statistical properties (e.g., mean) depend on the time when the data are collected.[20] Let us consider a concrete example (adapted from Hoover, 2003), which we call 'Height&Math':

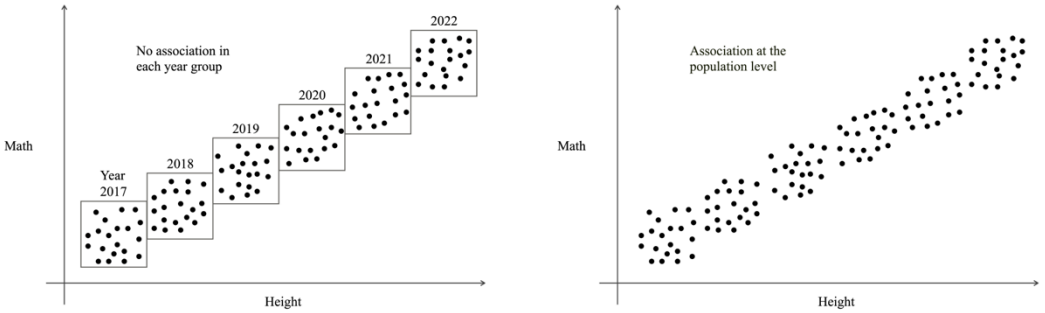
> *Height&Math*: Choose a class of 20 six-year-old children in the US and a class of 20 six-year-old children in China. Each year, we first measure the heights of the US children (*Height*) and order the data in the alphabetic order of their last names, and then measure the mathematical knowledge of the Chinese children (*Math*) using a standard diagnostic test, and order data in the same way.[21] Collect the data annually for six years.

In every single year, a US child's height and a Chinese child's mathematical knowledge are expected to be *unassociated*. The fact that a US child is of a certain height (e.g., 43 inches) should indicate nothing about what score the counterpart Chinese child earns on the math test, and vice versa. Yet, over the years, as the US children grow taller, the Chinese children also learn more math in school. As a result, in the data of all six years, both *Height* and *Math* will increase monotonically, producing a strong association between them. As illustrated in Figure 5, by treating the six years as a whole group and every single year as a sub-group, we see that an association

[20] Perhaps the most well-known example of this type is Sober's (2001) Venetian sea levels and British bread prices example (cf. Hoover, 2003; Steel, 2003). However, it is not easy to intuitively demonstrate why Sober's example is a case of SP: when we fix the year in which sea levels and bread prices are measured, the sample size reduces to 1, which makes it hard to show that the association between them disappears when conditioning on year.

[21] It is not a necessary feature of this example that the data be ordered by children's names. As long as we choose an ordering that is 'random' and use it consistently throughout the data collection process, SP can arise.

absent at the sub-group level (Figure 5a) emerges at the group level (Figure 5b), making this example a case of SP.

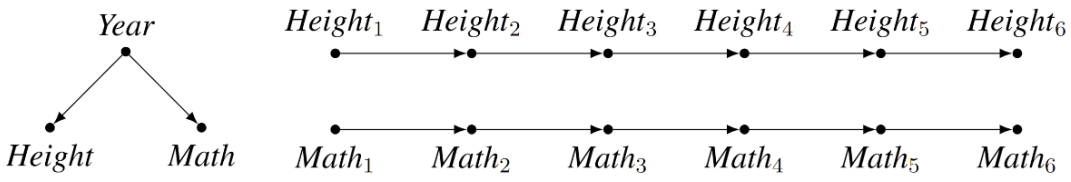


(a) When looking at data collected within each year (e.g., 2017), we observe no significant association

(b) When looking at the data collected over six years, we observe a strong association between *Math* and *Height*

Figure 5 An illustration of the Height&Math example.

Looking at this example, one might be tempted to posit *Year* as a common cause; doing so would allow this SP to be explained in terms of confounding. However, we believe this idea cannot stand scrutiny. Here, we agree with Yule (1926) and Steel (2003)—as well as J. Zhang and Spirtes (2014)—that it is ill-motivated to treat time as a cause. We are not saying this is uncontroversial, but given this broad consensus in the literature, the burden of proof is on those who think otherwise. The graph in Figure 6a is, therefore, not an appropriate representation of the causal structure behind the Height&Math example.



(a) A wrong DAG

(b) A more accurate DAG with two series of variables

Figure 6 Two DAGs for the Height&Math example.

We believe the best diagnosis of how SP arises in this example is as follows (see Spirtes et al., 2000, p. 296; J. Zhang & Spirtes, 2014). So far, we have used 'height' and 'mathematical knowledge' as two ordinary random variables, which are measured over several years on the chosen children. However, a child's height (or math knowledge) in one year is *causally* relevant to the child's height (or math knowledge) in the next year, because the child in one year grows (or learns) based on the last year's height (or math knowledge). The original unit choice and variable choice thus induce '*inter-unit*' causation, meaning that there is a causal relation between how one unit instantiates the relevant properties and how another unit instantiates them. In other words, the statistical units selected for analyzing a set of variables causally interact with each other. This creates trouble for standard statistical and causal analysis because inter-unit causation leads to the violation of the IID (independent and identically distributed) assumption in the sampling process. Causal analysis, built upon statistical analysis, is designed to capture *intra*-unit causation, not *inter*-unit causation (J. Zhang & Spirtes, 2014).

Given the above diagnosis, the solution is to redefine height and mathematical knowledge as two *series* of variables: $\mathbf{Height}_t = \{Height_1, \ldots, Height_6\}$ and $\mathbf{Math}_t = \{Math_1, \ldots, Math_6\}$, where each variable is indexed to a year. This new unit choice and variable choice induces no inter-unit causation, and also successfully captures the underlying causal story. The resulting DAG is shown in Figure 6b, which represents increases in height and mathematical knowledge using two parallel but separate causal chains. Comparing Figure 6b against Figure 6a makes it clear that "[i]t is the inter-unit causation that propagates an initial coincidence into [an] association", as J. Zhang and Spirtes (2014, p. 247) summarize.

In our opinion, we need a DAG at least as complex as the one in Figure 6b to accurately depict the causal structure behind the kind of SP involved in the Height&Math example, and in general, any SP due to inter-unit causation.[22] Note this causal analysis has gone far beyond what Pearl's analysis can offer. Rather than simply drawing a DAG among the three given variables (i.e., *Height*, *Math* and *Year*) involved in the case of SP in the Height&Math example, what the new analysis demands is a redefinition of the variables and a reselection of units such that inter-unit causation is avoided.

5. The amalgamation paradox and the non-collapsibility of odds ratio

All cases of SP discussed in the previous two sections share the common feature that the *X*-variable is associated with the partitioning variable *Z* (e.g., *Size* is associated with *Bag*; *Selfish* with *Group*, etc.). As mentioned earlier, the presence of this association is necessary for SP to occur. However, if we shift our focus from SP to its more generalized form, the *amalgamation paradox* (AP), the association between *X* and *Z* would no longer be necessary.[23]

AP is the statistical phenomenon in which the *marginal* or unconditional association of two variables, *X* and *Y*, falls outside the range of the *conditional* associations with respect to a third variable, *Z*. To illustrate this definition, consider an example from Greenland (2021). Suppose that a study on a medical treatment with high mortality has 50 males and 100 females in both the treatment and the control group. The two groups are relatively homogeneous in their medically relevant features such that no factor other than the treatment and sex can account for the difference

[22] Of course, this more complicated DAG is still a *simplified* representation of the true causal structure. For example, a child's current math knowledge is influenced by not only her math knowledge in the previous year, but also the amount of education she received during the year. Representing these extra causes will make the DAG more complete and more accurate, but will not affect our conclusion that time should be understood as what the variables are indexed to, rather than a causal variable on its own.

[23] As noted by Samuels (1993, p. 84), AP was first identified and defined by Good and Mittal (1987) who pointed out that SP implies AP: all cases of SP are cases of AP. See also Sprenger and Weinberger (2021) for a definition of AP and its relationship with SP.

in outcomes between the two groups. That is, in this example, *Treatment* is unassociated with *Sex*, and *Sex* is *not* a confounder (but an effect modifier; more below). Results show that among treated patients, 45 of 50 males and 30 of 100 females died, whereas, among untreated patients, 30 of 50 males and 10 of 100 females died (see Table 4).

	Males		**Females**		**Total**	
	Treated	*Untreated*	*Treated*	*Untreated*	*Treated*	*Untreated*
Totals	50	50	100	100	150	150
Died	45	30	30	10	75	40
Risks/Proportions	0.90	0.60	0.30	0.10	0.500	0.267
Risk differences	0.30		0.20		0.233	
Odds	9/1	3/2	3/7	1/9	1/1	4/11
Odds ratios	**6.0**		**3.9**		**2.75**	

Table 4 This table demonstrates a case of AP relative to the odds ratio (OR) measure: the marginal association between *Treatment* and *Death* in the total group is 2.75, which is smaller than conditional associations in both males and females (6.0 and 3.9 respectively). For comparison, AP does not arise when risk difference is the association measure.

The above results indicate that the medical treatment has effects of *different magnitudes* on mortality among males and females. This is true regardless of whether the effect measure is odds ratio (OR) or risk difference—both are widely used in epidemiology. But there is something counterintuitive about OR in the above example. When OR is the chosen measure, we have OR(*Treatment*, *Death*) < OR(*Treatment*, *Death* | *Sex* = Female) < OR(*Treatment*, *Death* | *Sex* = Male). This implies that the marginal association between *Treatment* and *Death* cannot be expressed as a weighted average of the two conditional associations with respect to *Sex*. That is, the marginal association falls outside the range of the conditional associations; this makes the example an instance of AP. Note that what gives rise to AP in the above example is the use of the

odds ratio measure, which has been known to be *non-collapsible* (Hernán et al., 2011).[24] By contrast, no such peculiarity is present when the chosen measure is risk difference, which is a *collapsible* measure, meaning that on this measure, the marginal association falls between the conditional associations.[25]

More broadly speaking, the peculiarity of AP lies in the fact that it violates our expectation that the association between two variables at the level of the entire group should be bounded by their conditional associations in the sub-groups (Hernán et al., 2011). This intuition comes from the seemingly plausible presupposition—which is not always true—that the marginal association should be a weighted average of the conditional associations. As a special case, SP violates a categorical or qualitative version of this intuition; namely, we expect that if two variables are conditionally unassociated or positively/negatively associated in each sub-group, they should likewise be unassociated or positively/negatively associated in the entire group. So, in cases of SP, not only the marginal association between *X* and *Y* falls outside the range of associations conditional on *Z*, but the marginal and the conditional associations are in *opposite* directions (e.g., the former is positive whereas the latter are non-positive). This is why SP, defined as association

[24] See also Cummings (2009) for more discussions on the non-collapsibility of odds ratios.

[25] The notion of (non-)collapsibility used here refers primarily to a property of an association measure. A measure, *m*, of the association between *X* and *Y* is collapsible across *Z*, if and only if the measured marginal association $m(X, Y)$ can be expressed as a *weighted average* of the measured conditional associations, $m(X, Y \mid Z)$ (cf. Pearl, 2009, p. 193; Huitfeldt et al., 2019). Risk difference, for instance, is a collapsible association measure in this sense. In contrast, odds ratio is a non-collapsible measure because there are datasets (such as the one represented in Table 4) where $\mathrm{OR}(X, Y)$ cannot be expressed as a weighted average of $\mathrm{OR}(X, Y \mid Z)$. One can also use collapsibility in a derivative sense that describe datasets. A dataset, D, is collapsible relative to a chosen association measure, *m*, if and only if the measured marginal association, $m(X, Y)$, reported in D can be expressed as a *weighted average* of the measured conditional associations, $m(X, Y \mid Z)$, in D. The dataset represented in Table 4 is non-collapsible relative to odds ratio in this sense. Moreover, Bandyopadhyay et al.'s (2015) notion of collapsibility states that a dataset is collapsible if and only if the marginal association between *X* and *Y* has the *same direction* as their conditional associations on *Z*. Since the marginal association being able to be expressed as a weighted average of the conditional associations implies that the marginal and the conditional associations are in the same direction, Bandyopadhyay et al.'s notion of (dataset) collapsibility encompasses the notion of (dataset) collapsibility as defined here in terms of weighted average. We thank an anonymous reviewer for prompting us to clarify this point.

reversal, disappearance, or emergence, is a special form of AP. In light of this, it is natural to extend our attention to AP and ask the following questions: how can AP arise in a causal context, and when it does, what should we do? Attempts at answering these questions will help delineate the scope of the application of Pearl's causal-graphical analysis of SP.

What is particularly noteworthy is that the kind of AP due to a choice of non-collapsible association measure typically occurs in the presence of effect modification.[26] Effect modification occurs when a causal effect is sensitive to the value of a third variable.[27] That is, *Z* is a *modifier* of the effect of *X* on *Y* when the effect varies across levels of *Z* (cf. Hernán & Robins, 2020). Importantly, a modifier is not a confounding variable. In our above example, *Sex* is a modifier of the treatment effect, but it is not a confounding variable since it is not associated with *Treatment*. This means that AP can occur in the absence of confounding. For this reason, Pearl's analysis of SP cannot be generalized to AP. To be fair, we are not claiming that this is an intrinsic difficulty with Pearl's analysis of SP, since to our knowledge, he does not claim that his analysis can be extended to dealing with AP. Still, we believe our discussion here helps us see the scope of Pearl's analysis.

Given that non-collapsibility is necessary for the type of AP discussed in this section, a general solution is to avoid using non-collapsible effect measures such as odds ratios without well-justified reasons. As for circumstances where we do want to use the odds ratio, we should be aware that compared with causal effect at the whole-group level, the effect of *X* on *Y* measured in the sub-groups stratified on the modifier *Z* conveys more accurate causal information about how *X*

[26] In other words, the combination of effect modification and non-collapsibility offers *an* important and typical causal context for the presence of AP in the data. However, this is not to say that effect modification is a necessary condition for the presence of AP. It is possible that the marginal odds ratio between *X* and *Y* differs from the conditional odds ratios when the latter two are equal. Nevertheless, the occurrence of AP without effect modification is much less frequent in real life, as well as less impressive, compared to cases in which AP is present due to effect modification.

[27] It is not easy to represent the presence of effect modification using standard DAGs; but see Weinberg (2007) for attempts of clarifying effect modification using (nonstandard) DAGs.

influences *Y* in a specific causal background. Thus, when the data display a pattern of AP due to non-collapsibility in the presence of effect modification, we should report sub-group-specific effect estimations (although this does not imply that effect estimation at the whole-group level is biased or meaningless).

6. Conclusion

In this article, we have argued that Pearl's causal-graphical analysis of SP, which diagnoses SP as essentially a peculiar consequence of confounding, cannot capture the full spectrum of the phenomenon. We show that there are good reasons to believe that SP is a generic term encompassing a wide range of distinct phenomena. Confounding is by no means the only source of SP, even if we admit that it is probably the most common and important one. Importantly, we do not claim that we have identified all possible sources of SP. There is nothing surprising about this, given that spurious associations observed in statistical data may come from a variety of sources: confounding, inappropriate variable aggregation, inter-unit causation, or sheer chance. The multiplicity of the sources of spurious association necessitates the multiplicity of the sources of SP. Thus, contra Pearl et al. (2016), we find it untenable that a confounding-based analysis can resolve all cases of SP. As far as we can see, the plurality of the sources of SP, and thereby the plurality of its resolutions, are here to stay.

References

Bandyopadhyay, P. S., Greenwood, M., Dcruz, D. W. F., & Raghavan R, V. (2015). Simpson's paradox and causality. *American Philosophical Quarterly*, 13-25.

Bandyopadhyay, P. S., Nelson, D., Greenwood, M., Brittan, G., & Berwald, J. (2011). The logic of Simpson's paradox. *Synthese*, *181*, 185-208.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, *187*(4175), 398-404.

Cartwright, N. (1979). Causal laws and effective strategies. *Noûs*, 419-437.

Cartwright, N. (2001). What is wrong with Bayes nets? *The Monist*, *84*(2), 242-264.

Cummings, P. (2009). The relative merits of risk ratios and odds ratios. *Arch Pediatr Adolesc Med* *163*(5), 438-45.

Dawid, A.P. (2010). Beware of the DAG!. *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, in *Proceedings of Machine Learning Research* 6:59-86 Available from https://proceedings.mlr.press/v6/dawid10a.html.

Didelez, V., & Stensrud, M. J. (2022). On the logic of collapsibility for causal effect measures. *Biometrical Journal*, 64(2), 235-242.

Dobzhansky, T., & Pavlovsky, O. (1957). An experimental study of interaction between genetic drift and natural selection. *Evolution*, 311-319.

Frisch, M. (2020). Causation in physics. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/win2023/entries/causation-physics/.

Good, I. J., & Mittal, Y. (1987). The amalgamation and geometry of two-by-two contingency tables. *The Annals of Statistics*, 694-711.

Greenland, S. (2021). Noncollapsibility, confounding, and sparse-data bias. Part 2: What should researchers make of persistent controversies about the odds ratio? *Journal of Clinical Epidemiology 139,* 264-268.

Hernán, M. A., Clayton, D., & Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology*, *40*(3), 780-785.

Hernán, M. A. & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC.

Hoover, K. D. (2003). Nonstationary time series, cointegration, and the principle of the common cause. *British Journal for the Philosophy of Science*, *54*(4), 527-551.

Huitfeldt, A., Stensrud, M. J., & Suzuki, E. (2019). On the collapsibility of measures of effect in the counterfactual causal framework. *Emerging themes in epidemiology*, 16, 1-5.

Irzik, G., & Meyer, E. (1987). Causal modeling: New directions for statistical explanation. *Philosophy of Science, 54*(4), 495-514.

Mukherjee, S. (2015). *The laws of medicine: field notes from an uncertain science*. Simon and Schuster.

Myrvold, W., Genovese, M., & Shimony, A. (2024). Bell's theorem. *The Stanford Encyclopedia of Philosophy*, E. N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/spr2024/entries/bell-theorem/.

Näger, P. M. (2022). Evidence for interactive common causes. Resuming the Cartwright-Hausman-Woodward debate. *European Journal for Philosophy of Science 12*, Article 2.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pearl, J. (2014). Comment: Understanding Simpson's paradox. *The American Statistician*, 68(1), 8-13.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic books.

Samuels, M. L. (1993). Simpson's paradox and related phenomena. *Journal of the American Statistical Association*, *88*(421), 81-88.

Sarkar, P., & Bandyopadhyay, P. S. (2021). Simpson's paradox: A singularity of statistical and inductive inference. *arXiv preprint arXiv:2103.16860*.

Shrier, I., & Platt, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, *8*(1), 1-15.

Sober, E. (2001). Venetian sea levels, British bread prices, and the principle of the common cause. *British Journal for the Philosophy of Science*, *52*(2), 331-346.

Sober, E. (2024). *The philosophy of evolutionary theory: Concepts, inferences, and probabilities*. Cambridge University Press.

Spanos, A. (2021). Yule–Simpson's paradox: The probabilistic versus the empirical conundrum. *Statistical Methods & Applications*, *30*, 605-635.

Sprenger, J. & Weinberger, N. (2021). Simpson's paradox. *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (ed.), https://plato.stanford.edu/archives/sum2021/entries/paradox-simpson/.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT Press.

Spirtes, P., & Scheines, R. (2004). Causal inference of ambiguous manipulations. *Philosophy of Science*, *71*(5), 833-845.

Steel, D. (2003). Making time stand still: A response to Sober's counter-example to the principle of the common cause. *British Journal for the Philosophy of Science*, *54*(2), 309-318.

Weinberg, C. R. (2007). Can DAGs clarify effect modification? *Epidemiology*, *18*(5), 569-572.

Wood, C. J., & Spekkens, R. W. (2015). The lesson of causal discovery algorithms for quantum correlations: Causal explanations of Bell-inequality violations require fine-tuning. *New Journal of Physics*, *17*(3), Article 033002.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Woodward, J. (2016). The problem of variable choice. *Synthese*, *193*, 1047-1072.

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series?--a study in sampling and the nature of time-series. *Journal of the Royal Statistical Society*, *89*(1), 1-63.

Zhang, J. & Spirtes, P. (2014). Choice of units and the causal Markov condition. In G. Guo & C. Liu (Eds.), *Scientific explanation and methodology of science: Selected essays from the international conference on SEMS 2012* (pp. 240-251). World Scientific.

Zhang, C., Mohan, K., & Pearl, J. (2022). Causal inference with non-IID data using linear graphical models. *Advances in Neural Information Processing Systems*, *35*, 13214-13225.