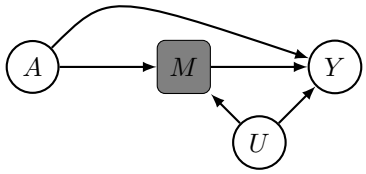
eAppendix "Re. Collider bias is only a partial explanation for the obesity paradox"

1 The causal model considered by Sperrin et al.

We refer to[2,3,4] for a thorough introduction of causal models and counterfactuals. Here, the fundamental concepts are illustrated in the particular case of the model considered in[1]. Figure 1 presents the Directed Acyclic Graph (DAG) attached to this causal model. In this DAG, only the endogenous variables U, A, M, Y are represented, while the corresponding exogenous variables, or disturbances, ε_U, ε_A, ε_M and ε_Y are not. The relationships between the endogenous variables can be fully described by the set of structural equations corresponding to this DAG. There is one such equation for each endogenous variable involved in the DAG. It involves a fixed (but unknown) autonomous function, whose inputs are the parents of the variable in the DAG, along with its associated exogenous variable. In our example, the set of structural equations is

$$\begin{cases} A &= f_A(\varepsilon_A) \\ U &= f_U(\varepsilon_U) \\ M &= f_M(A, U, \varepsilon_M) \\ Y &= f_Y(A, M, U, \varepsilon_Y), \end{cases}$$

Figure 1: The DAG considered in [1]



where, f_A, f_U, f_M and f_Y are the unspecified autonomous functions. The exogenous variables are usually assumed to be mutually independent, in which case the causal model is said to be Markov; keep in mind that the associated graph is a DAG here [2].

The structural equations are further helpful to precisely define the counterfactual variables, or potential outcomes, which correspond to the variables we would have been able to observe had we intervened to fix the value(s) of some variable(s). Consider for instance the counterfactual world $\Omega^{A=a}$ that would have followed the intervention $do(A = a)$ [2,5,6]. We may mention that such an intervention is not well-defined when A represents obesity. As nicely put forward in [7], this makes causal inference about obesity a particularly difficult task, because assumptions such as consistency, positivity and exchangeability are unlikely to hold [7]. Here we ignore this problem just as Sperrin et al. did. In other words, we assume the existence of such an intervention and proceed as usual under structural causal models.

After the intervention $do(A = a)$, we would have been able to observe the variables

$$\begin{cases} U & = f_U(\varepsilon_U) \\ M^{A=a} & = f_M(a, U, \varepsilon_M) \\ Y^{A=a} & = f_Y(a, M^{A=a}, U, \varepsilon_Y). \end{cases}$$

From these equations, it is clear that consistency holds: if $A = a$, then $M = f_M(a, U, \varepsilon_M) = M^{A=a}$ and $Y = f_Y(a, M, U, \varepsilon_Y) = f_Y(a, M^{A=a}, U, \varepsilon_Y) = Y^{A=a}$.

For future use, we may recall that counterfactual variables are random variables in the usual sense, even if they are not fully observed. Indeed, the counterfactual variables and observed variables are all defined on a common probability space as deterministic functions of the exogenous variables ε_A, ε_M, ε_U and ε_Y; see Sections 4 and 5 in[5] for instance. As a result, standard probability calculus applies on counterfactual variables. For instance, the tower rule, which states that $E(Z) = E[E(Z|X)]$ for "any" couple of random variables X and Z, can be applied with, say, $X = Y^{A=a}$ and $Z = M$, leading to $P[Y^{A=a} = 1] = \sum_m P[Y^{A=a} = 1|M = m]P[M = m]$.

In other respect, the variables that we would have been able to observe in the counterfactual world $\Omega^{a,m}$ that would have followed the double intervention $do(A = a, M = m)$ can be defined similarly:

$$\begin{cases} U & = f_U(\varepsilon_U) \\ Y^{A=a,M=m} & = f_Y(a, m, U, \varepsilon_Y). \end{cases}$$

From these definitions, we have $Y^{A=a} = Y^{A=a,M=M^{A=a}}$. Here $Y^{A=a,M=M^{A=a}}$ can be thought of as the outcome we would have been able to observe in the counterfactual world that would have followed the intervention $A = a$ and $M = M^{A=a}$, that is the counterfactual world where A would have been set to a and M would have been set to whatever values it can get under the distribution of $M^{A=a}$; of course this counterfactual world is exactly $\Omega^{A=a}$.

2 Controlled direct effects

The quantity $\Delta_{CE} = P[Y^{A=1} = 1|M = 1] - P[Y^{A=0} = 1|M = 1]$ has to be interpreted with caution. Because $Y^{A=a} = Y^{A=a,M=M^{A=a}}$, it comes

$$P[Y^{A=a} = 1|M = 1] = P[Y^{A=a,M=M^{A=a}} = 1|M = 1].$$

In particular, $P[Y^{A=0} = 1|M = 1] = P[Y^{A=0,M=M^{A=0}} = 1|M = 1]$ is the risk of early death in $\Omega^{A=0}$, for the individuals with diabetes in the actual world. But $M = 1 \not\Rightarrow M^{A=0} = 1$: some diabetic patients in the actual world would not have been diabetic in $\Omega^{A=0}$. As a result, a positive Δ_{CE}, for instance, can be solely due to the fact that a portion of the individuals with diabetes in the actual world would not have suffered from diabetes in $\Omega^{A=0}$ and would have lived longer in $\Omega^{A=0}$ than they would have in $\Omega^{A=1}$. In other words, even if it is conditioned on $M = 1$, Δ_{CE} is not related to the direct effect of A on Y. It is the total effect of A among one particular subgroup of the population defined according to a variable observed in the actual world, just as the average causal effect on the treated $E[Y^{A=1} - Y^{A=0}|A = 1]$ is a measure of the total effect of A in another subgroup of the population.

In the presence of a mediator like M here, alternative causal effects have been advocated[8,9] as measures of the direct effect. In particular, the controlled direct effect captures the effect of A, while fixing the value of the mediator (*e.g.*, to 1), and is therefore appealing in the context of the obesity paradox; see Section 5 below for more details. For simplicity, we set $Y^{a,m} = Y^{A=a,M=m}$. The three following quantities can be considered:

$$\begin{aligned}
\Delta_{CDE} &= P[Y^{1,1} = 1] - P[Y^{0,1} = 1] \\
\Delta_{CDE|M=1} &= P[Y^{1,1} = 1|M = 1] - P[Y^{0,1} = 1|M = 1] \\
\Delta_{CDE|A=1,M=1} &= P[Y^{1,1} = 1|A = 1, M = 1] - P[Y^{0,1} = 1|A = 1, M = 1].
\end{aligned}$$

The first one, Δ_{CDE}, is the controlled direct effect of A on Y, at $M = 1$[8,9]. It compares the risk of death in counterfactual worlds $\Omega^{1,1}$ and $\Omega^{0,1}$, that would have followed the double interventions $do(A = 1, M = 1)$ and $do(A = 0, M = 1)$, respectively. All individuals suffer from diabetes in these two counterfactual worlds, but they are all obese in $\Omega^{1,1}$, while none of them is obese in $\Omega^{0,1}$. Therefore, Δ_{CDE} captures the direct causal effect of obesity, while controlling for the diabetic status; $M = 1$ here. The

other two quantities are conditional versions of Δ_{CDE}. More precisely, $\Delta_{CDE|M=1}$ corresponds to Δ_{CDE} when focusing on individuals who suffer from diabetes in the actual world, while $\Delta_{CDE|M=1,A=1}$ focuses on individuals who are obese and suffer from diabetes in the actual world.

Under the structural causal model corresponding to Figure 1, both Δ_{CDE} and $\Delta_{CDE|A=1,M=1}$ can be identified from the distribution of (A, M, U, Y). Moreover, the corresponding formulas exhibit similarities with both Δ_{AS} and Δ_{Sp}. Because there is no unobserved confounder except U between A and Y, we have $Y^{a,m} \perp\!\!\!\perp A|U$.[9] Similarly, because there is no unobserved confounder except U between M and Y, we have and $Y^{a,m} \perp\!\!\!\perp M|(A, U)$.[9] Therefore,

$$
\begin{aligned}
P[Y^{a,m}=1] &= \sum_u P[Y^{a,m}=1|U=u]P[U=u] \quad \text{by the tower rule} \\
&= \sum_u P[Y^{a,m}=1|A=a,U=u]P[U=u] \quad \text{because } Y^{a,m} \perp\!\!\!\perp A|U \\
&= \sum_u P[Y^{a,m}=1|A=a,M=m,U=u]P[U=u] \quad \text{because } Y^{a,m} \perp\!\!\!\perp M|(A,U) \\
&= \sum_u P[Y=1|A=a,M=m,U=u]P[U=u] \quad \text{by consistency.}
\end{aligned}
$$

Therefore, Δ_{CDE} can be written

$$
\begin{aligned}
\Delta_{CDE} = \sum_u \{&P[Y=1|A=1,M=1,U=u] \\
&- P[Y=1|A=0,M=1,U=u]\}P(U=u). \qquad (1)
\end{aligned}
$$

Now, turning our attention to $\Delta_{CDE|A=1,M=1}$, we have, for any $(a_1, a_2, m_1, m_2) \in$

$\{0,1\}^4$,

$$
\begin{aligned}
&P[Y^{a_1,m_1}=1|A=a_2,M=m_2] \\
&=\sum_u P[Y^{a_1,m_1}=1|A=a_2,M=m_2,U=u]P(U=u|A=a_2,M=m_2) \quad \text{by the tower rule} \\
&=\sum_u P[Y^{a_1,m_1}=1|A=a_2,U=u]P(U=u|A=a_2,M=m_2) \quad \text{because } Y^{a,m} \perp\!\!\!\perp M|(A,U) \\
&=\sum_u P[Y^{a_1,m_1}=1|A=a_1,U=u]P(U=u|A=a_2,M=m_2) \quad \text{because } Y^{a,m} \perp\!\!\!\perp A|U \\
&=\sum_u P[Y^{a_1,m_1}=1|A=a_1,M=m_1,U=u]P(U=u|A=a_2,M=m_2) \quad \text{because } Y^{a,m} \perp\!\!\!\perp M|(A,U) \\
&=\sum_u P[Y=1|A=a_1,M=m_1,U=u]P(U=u|A=a_2,M=m_2) \quad \text{by consistency.}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\Delta_{CDE|A=1,M=1} = \sum_u \{&P[Y=1|A=1,M=1,U=u] \\
&- P[Y=1|A=0,M=1,U=u]\}P(U=u|A=1,M=1). \quad (2)
\end{aligned}
$$

Turning our attention back on Δ_{AS}, observe that it writes

$$
\begin{aligned}
\sum_u \{&P[Y=1|A=1,M=1,U=u]P(U=u|A=1,M=1) \\
&-P[Y=1|A=0,M=1,U=u]P(U=u|A=0,M=1)\}.
\end{aligned}
$$

Sperrin et al. claim that the difference between Δ_{AS} and Δ_{CE} is possible because $P(U=u|A=a,M=m)$ generally differs from $P(U=u|M=m)$. As we showed in our Letter, their claim is only valid for the difference between Δ_{AS} and Δ_{Sp}, and the difference between Δ_{AS} and Δ_{CE} is mostly due to the discrepancy between $P(Y=1|A=a,M=1)$ and $P[Y^{A=a}|M=1]$, which is precisely caused by collider bias. From the above formula, Sperrin et al.'s discussion is actually valid

for the difference between Δ_{AS}, Δ_{Sp}, Δ_{CDE} and $\Delta_{CDE|A=1,M=1}$. Indeed, the only differences between these four quantities lie in the version of the U-distribution used to marginalize the quantities $P[Y=1|A=a,M=1,U=u]$ over u, for $a \in \{0,1\}$. It is $P(U=u|A=a,M=1)$ for Δ_{AS}, $P(U=u|M=1)$ for Δ_{Sp}, $P(U=u)$ for Δ_{CDE} and $P(U=u|A=1,M=1)$ for $\Delta_{CDE|A=1,M=1}$.

Despite these similarities, these four quantities are generally different since M typically depends on both A and U under the model of Figure 1. In particular, the bias between Δ_{AS} and Δ_{CDE}, or between Δ_{AS} and $\Delta_{CDE|A=1,M=1}$, is due to confounding. Even if the numerical results presented in our Letter show that this bias is typically much lower than that between Δ_{AS} and Δ_{CE}, we may mention that it is still possible to have $\Delta_{AS} < 0$ while $\Delta_{CDE} > 0$ and $\Delta_{CDE|A=1,M=1} > 0$; see Section 3.3 below.

Way may further mention that quantities $\Delta_{CDE|M=1}$ and Δ_{Sp} are generally different. These two quantities would be equal if $Y^{a,m} \perp\!\!\!\perp A|(M,U)$, but this conditional independence does usually not hold under the model of Figure 1. More generally, we were not able to relate Δ_{Sp} to any meaningful causal effect under this model.

A final remark is that $\Delta_{CDE|M=1}$ can not be identified from the distribution of (A,M,U,Y) under the model of Figure 1 without further assumptions. It may be identifiable after specifying the generating functions and the disturbances distributions following the same arguments as those used in the case of Δ_{CE}; see Section 3.2 below.

3 Numerical illustration

3.1 Data generation mechanism

Here we describe the generative model that we used to illustrate the differences between the various quantities considered in our Letter and in the present Web Appendix. We consider a generative model that can be seen as a special case of, and is then consistent with, the one described by[1]. More precisely, our data generation mechanism

is obtained by specifying the structural functions f_Y, f_M, f_A and f_U as well as the distributions of the disturbances ε_Y, ε_M, ε_A and ε_U, which together lead to the same relationships between Y, A, M and U as those considered in Sperrin et al. Keep in mind that the causal model has to be specified in order to derive an analytic formula for Δ_{CE} (see Section 3.2 below).

Denote the indicator function by $\mathbb{1}[\cdot]$. Define four independent random variables ε_A, ε_U, ε_M and ε_Y distributed according to a uniform distribution over the interval $[0, 1]$. For any given $(p_A, p_U) \in (0, 1)^2$, define $A = \mathbb{1}[\varepsilon_A \leq p_A]$ and $U = \mathbb{1}[\varepsilon_U \leq p_U]$ so that $A \sim B(p_A)$ and $U \sim B(p_U)$ are two independent Bernoulli variables. As in Sperrin et al.'s work, we consider the special case where $p_A = p_U = 0.5$. Now, introduce the sigmoid function $\text{expit}(x) = (1 + \exp(-x))^{-1}$, and set, for any $(a, u, m) \in \{0, 1\}^3$ and for some real parameters $\alpha_0, \alpha_A, \alpha_U, \alpha_{AU}, \beta_0, \beta_A, \beta_U, \beta_M, \beta_{AM}, \beta_{AU}, \beta_{UM}$ and β_{AUM},

$$
\begin{aligned}
p_M(a, u) &= \text{expit}(\alpha_0 + \alpha_A a + \alpha_U u + \alpha_{AU} au) \\
p_Y(a, m, u) &= \text{expit}(\beta_0 + \beta_A a + \beta_U u + \beta_M m + \beta_{AU} au + \\
&\qquad \beta_{AM} am + \beta_{UM} um + \beta_{AUM} aum).
\end{aligned}
$$

Finally, variables M and Y are defined as

$$
\begin{aligned}
M &= f_M(A, U, \varepsilon_M) = \mathbb{1}[\varepsilon_M \leq p_M(A, U)], \\
Y &= f_Y(A, M, U, \varepsilon_Y) = \mathbb{1}[\varepsilon_Y \leq p_Y(A, M, U)].
\end{aligned}
$$

Sperrin et al. only considered situations where interaction terms in the Y-model were absent: $\beta_{AM} = \beta_{AU} = \beta_{UM} = \beta_{AUM} = 0$. We will show below that conclusions can be quite different when considering non-zero values for these parameters, especially when comparing Δ_{AS} and Δ_{CDE}. Moreover, following Sperrin et al., we

set

$$\alpha_0 = -\frac{1}{2}\left(\alpha_A + \alpha_U + \frac{1}{2}\alpha_{AU}\right)$$
$$\beta_0 = -\frac{1}{2}\left(\beta_A + \beta_M + \beta_U + \frac{1}{2}(\beta_{AM} + \beta_{AU} + \beta_{UM}) + \frac{1}{4}\beta_{AUM}\right).$$

3.2 Analytic formula of Δ_{CE} under our generative model

Recall that $p_A = p_U = 1/2$. First,

$$\begin{aligned}
&P(Y^{A=a} = 1|M = 1)\\
&= \sum_{\substack{i_A\in\{0,1\}\\ i_U\in\{0,1\}}} P(Y^{A=a} = 1|M = 1, A = i_A, U = i_U)P(A = i_A, U = i_U|M = 1)\\
&= \sum_{\substack{i_A\in\{0,1\}\\ i_U\in\{0,1\}}} P(Y^{A=a} = 1, M = 1|A = i_A, U = i_U)\frac{P(A = i_A, U = i_U|M = 1)}{P(M = 1|A = i_A, U = i_U)}\\
&= \sum_{\substack{i_A\in\{0,1\}\\ i_U\in\{0,1\}}} P(Y^{A=a} = 1, M = 1|A = i_A, U = i_U)\frac{P(A = i_A, U = i_U)}{P(M = 1)}\\
&= \sum_{\substack{i_A\in\{0,1\}\\ i_U\in\{0,1\}}} P(Y^{A=a} = 1, M = 1|A = i_A, U = i_U)\frac{1}{4P(M = 1)}\\
&\overset{(*)}{=} \sum_{\substack{i_A\in\{0,1\}\\ i_U\in\{0,1\}}} \int_0^{p_M(i_A,i_U)} \frac{P(Y^{A=a} = 1|\varepsilon_M = \varepsilon, A = i_A, U = i_U)}{4P(M = 1)}d\varepsilon\\
&= \sum_{\substack{i_A\in\{0,1\}\\ i_U\in\{0,1\}}} \int_0^{p_M(i_A,i_U)} \frac{P(\varepsilon_Y \le p_Y(a, M^{A=a}, U)|\varepsilon_M = \varepsilon, A = i_A, U = i_U)}{4P(M = 1)}d\varepsilon
\end{aligned}$$

where we use the facts that (i) $M = 1|(A = i_A, U = i_U)$ is equivalent to $\varepsilon_M \le p_M(i_A, i_U)$ and (ii) the conditional density of ε_M given $(A = i_A, U = i_U)$ uniformly equals 1 over the interval $[0, 1]$ to establish equality $(*)$. Then, successively using the fact that (i) $M^{A=a} = \mathbb{1}[\varepsilon_M \le p_M(a, U)]$, and (ii) $P(\varepsilon_Y \le \rho) = \rho$ for any $\rho \in [0, 1]$,

and setting $x \wedge y = \min(x, y)$, it follows that

$$
\begin{aligned}
&P(Y^{A=a} = 1 | M = 1) \\
&= \sum_{\substack{i_A \in \{0,1\} \\ i_U \in \{0,1\}}} \frac{\int_0^{[p_M(i_A,i_U) \wedge p_M(a,i_U)]} P(\varepsilon_Y \leq p_Y(a,1,i_U)) d\varepsilon + \int_{[p_M(i_A,i_U) \wedge p_M(a,i_U)]}^{p_M(i_A,i_U)} P(\varepsilon_Y \leq p_Y(a,0,i_U)) d\varepsilon}{4P(M=1)} \\
&= \sum_{\substack{i_A \in \{0,1\} \\ i_U \in \{0,1\}}} \frac{\int_0^{[p_M(i_A,i_U) \wedge p_M(a,i_U)]} p_Y(a,1,i_U) d\varepsilon + \int_{[p_M(i_A,i_U) \wedge p_M(a,i_U)]}^{p_M(i_A,i_U)} p_Y(a,0,i_U) d\varepsilon}{4P(M=1)} \\
&= \sum_{\substack{i_A \in \{0,1\} \\ i_U \in \{0,1\}}} \frac{p_Y(a,1,i_U)[p_M(i_A,i_U) \wedge p_M(a,i_U)] + p_Y(a,0,i_U)\{p_M(i_A,i_U) - [p_M(i_A,i_U) \wedge p_M(a,i_U)]\}}{4P(M=1)}.
\end{aligned}
$$

Because $P(M = 1) = \sum_{i_A, i_U} p_M(i_A, i_U)/4$, it is straightforward to compute $P(Y^{A=a} = 1 | M = 1)$, hence Δ_{CE}, for any combinations of values for the parameters $\alpha_A, \alpha_U, \alpha_{AU}, \beta_A, \beta_U, \beta_M, \beta_{AM}, \beta_{AU}, \beta_{UM}$ and β_{AUM}.

3.3 Results

To be consistent with Sperrin et al.'s article, we present measures of association and causal effects on the odds-ratio scale, rather than on the difference scale. The corresponding quantities are denoted by OR_{AS}, OR_{Sp}, OR_{CE}, OR_{CDE} and $OR_{CDE|A=1,M=1}$. For instance,

$$
\begin{aligned}
OR_{AS} &= \frac{P(Y=1|A=1,M=1)/P(Y=0|A=1,M=1)}{P(Y=1|A=0,M=1)/P(Y=0|A=0,M=1)} \\
OR_{CE} &= \frac{P[Y^{A=1}=1|M=1]/P[Y^{A=1}=0|M=1]}{P[Y^{A=0}=1|M=1]/P[Y^{A=0}=0|M=1]} \\
OR_{CDE} &= \frac{P[Y^{1,1}=1]/P[Y^{1,1}=0]}{P[Y^{0,1}=1]/P[Y^{0,1}=0]} \\
OR_{CDE|A=1,M=1} &= \frac{P[Y^{1,1}=1|A=1,M=1]/P[Y^{1,1}=0|A=1,M=1]}{P[Y^{0,1}=1|A=1,M=1]/P[Y^{0,1}=0|A=1,M=1]}.
\end{aligned}
$$

We first consider the setting corresponding to Figure 2 of Sperrin et al., where

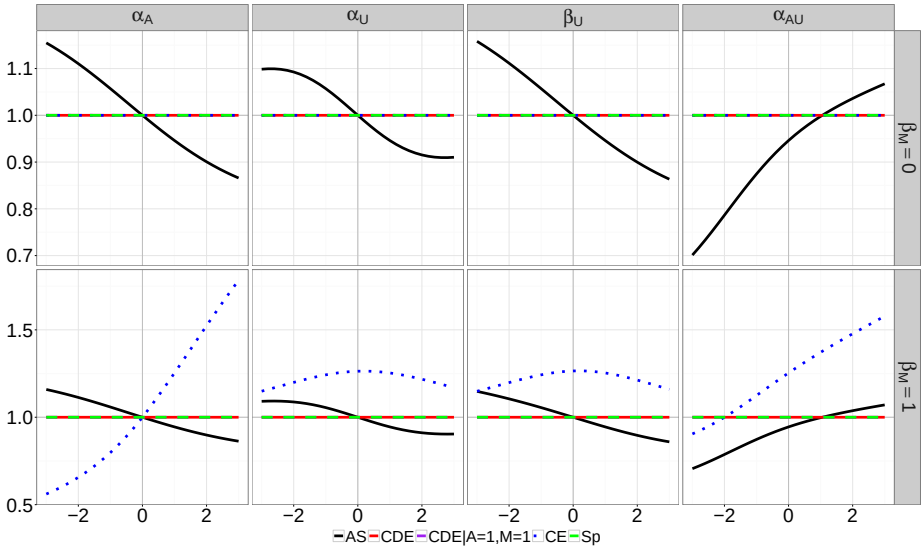
$\beta_A = \beta_{AM} = \beta_{AU} = \beta_{UM} = \beta_{AUM} = 0$ and $\beta_M = 0$; see the top row of Figure 2, which corresponds to the top row of Figure 2 in our Letter. We obtain the exact same results as Sperrin et al. Indeed, this setting corresponds to the case where A has neither a direct nor an indirect effect, and causal odds-ratios OR_{CE}, OR_{CDE} and $OR_{CDE|A=1,M=1}$ all equal 1. This is actually one particular situation where $Y^{A=a} \perp\!\!\!\perp A|M$ under the model of Figure 1; see Section 4.3 below for some additional remarks under a simplified model. Therefore $OR_{Sp} = OR_{CE}$ =1. On the other hand, OR_{AS} is generally not equal to 1, but the difference with the other quantities is typically small.

According to Sperrin al., β_M can be set to 0 without loss of generality. The bottom row of Figure 2 shows that it is not the case just as the bottom two rows of Figure 2 in our Letter. Indeed, when $\beta_M \neq 0$, A has an indirect effect, hence a total effect, and OR_{CE} is typically different from 1, as discussed in Section 2 above. On the other hand, the quantity OR_{Sp} still equals 1, and so do OR_{CDE} and $OR_{CDE|A=1,M=1}$. As for OR_{AS}, it behaves as in the case where $\beta_M = 0$. This particular case illustrates the discrepancy between the true value of OR_{CE} and the quantity OR_{Sp} studied by Sperrin et al. It further shows that if Δ_{CE} is the target quantity, then OR_{AS} can be severely biased and, then, that Sperrin et al.'s conclusion is false. In particular when $\alpha_A > 2$ and the other parameters are set to their default values, OR_{AS} is sensibly lower than 1 while OR_{CE} is sensibly greater than 1.

However, if the target quantity is OR_{CDE} or $OR_{CDE|A=1,M=1}$ the bias attached to OR_{AS} is less sensible under the configurations presented on Figure 2. We present results under configurations where $\alpha_A = \alpha_U = \beta_A = \beta_{UM} = 2$, $\beta_U = 3$ and $\beta_{AM} = -2$. In each panel of Figure 3, one of the four remaining parameters, α_{AU}, β_M, β_{AU} and β_{AUM}, varies between -3 and 3, while the other three are fixed at 1. Overall, under these configurations, OR_{AS} is sensibly inferior to 1 while OR_{CDE}, $OR_{CDE|A=1,M=1}$ and OR_{CE} are all sensibly superior to 1.

To recap, our numerical results establish that a negative association between A and

Figure 2: Causal and "observable"odds-ratios in the case where $\beta_A = \beta_{AM} = \beta_{AU} = \beta_{UM} = \beta_{AUM} = 0$, with β_M set to either 0 or 1, and for varying values of the other parameters $\alpha_A, \alpha_U, \beta_U$ and α_{AU} . In each panel, along the x axis, one of these parameters is varied from -3 to 3 (left panel: α_A, mid-left panel: α_U, mid-right panel: β_U, right panel: α_{AU}), and the other parameters are set to a default value (1 for $\alpha_A, \alpha_U, \beta_U$ and 0 for α_{AU}).



Y, when restricting our attention to patients with $M = 1$, $OR_{AS} < 1$, does not imply either $OR_{CE} < 1$, or $OR_{CDE} < 1$ or $OR_{CDE|A=1,M=1} < 1$. Therefore, even under the simple generative model considered by Sperrin et al., the "obesity paradox" can be artifactual and fully due to collider or confounding bias.

4 Additional remarks under a simplified model

4.1 The simplified model

It is instructive to inspect the simplified causal model of Figure 4, which is a special case of that of Figure 1. As such, a result that is generally false under this simplified causal model is generally false under the model considered by Sperrin et al. too.

In this simplified model, A is not a parent of Y, and there is no confounder. Then,

Figure 3: Causal and "observable"odds-ratios in the case where $\alpha_A = \alpha_U = 2 = \beta_A = \beta_{UM} = 2$, $\beta_U = 3$ and $\beta_{AM} = -2$, for varying values of the other parameters α_{AU}, β_M, β_{AU} and β_{AUM}. In each panel, along the x axis, one of these parameters is varied from -3 to 3 (left panel: α_{AU}, mid-left panel: β_M, mid-right panel: β_{AU}, right panel: β_{AUM}) and the other parameters are set to the default value 1.

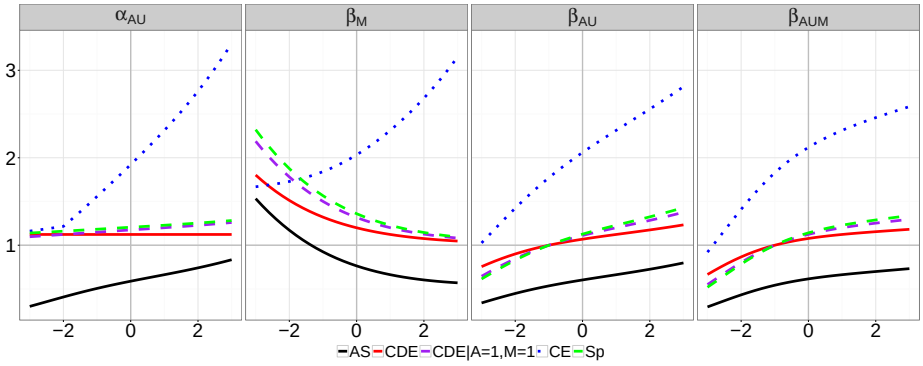
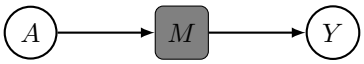


Figure 4: A simplified version of the DAG considered in Sperrin et al., corresponding to the special case of no confounder and no direct effect of A on Y.



the set of structural equations becomes

$$\begin{cases} A &= f_A(\varepsilon_A) \\ M &= f_M(A, \varepsilon_M) \\ Y &= f_Y(M, \varepsilon_Y). \end{cases}$$

Accordingly, the counterfactual variables $Y^{A=a}$, $M^{A=a}$ and $Y^{a,m}$ are defined as

$$\begin{cases} M^{A=a} & = f_M(a, \varepsilon_M) \\ Y^{A=a} & = f_Y(M^{A=a}, \varepsilon_Y) \\ Y^{a,m} & = f_Y(m, \varepsilon_Y). \end{cases}$$

4.2 Total effects, controlled direct effects and Δ_{CE}

In this DAG, the empty set satisfies the back-door criterion and, then, $Y^{A=a} \perp\!\!\!\perp A$. As a result, $P[Y^{A=a} = 1] = P[Y = 1|A = a]$, and the average total effect is generally non-null, as expected:

$$P[Y^{A=1} = 1] - P[Y^{A=0} = 1] = P[Y = 1|A = 1] - P[Y = 1|A = 0] \neq 0. \quad (3)$$

More precisely, this total effect is always non-null except in the absence of either the arrow pointing from A to M or the arrow pointing from M to Y.

Next, we have $Y \perp\!\!\!\perp A|M$ so that $P[Y = 1|A = 1, M = 1] = P[Y = 1|A = 0, M = 1]$ and $\Delta_{AS} = 0$. Given the expression (1) and (2), it follows that $\Delta_{CDE} = 0$ and $\Delta_{CDE|A=1,M=1} = 0$ too.

However, M is still a collider here, its parents being A and ε_M. Then, conditioning on M typically creates spurious correlation between ε_A and ε_Y. Therefore, collider bias is still at play, provided A has a causal effect on M and M has a causal effect on Y. As a matter of fact, $\{M\}$ does not satisfy the back-door criterion, and $Y^{A=a} \perp\!\!\!\perp A|M$ generally does not hold: $P[Y^{A=a} = 1|M = 1]$ generally differs from $P[Y = 1|A = a, M = 1]$. Consequently, Δ_{CE} is generally non-null, contrary to what Sperrin et al. wrote on the right column of Page 526. Indeed, considering the case where $\beta_A = \beta_{AM} = 0$ under their generative model, which corresponds to a situation where A has no direct effect on Y, they claim that $P[Y^{A=1} = 1|M = 1] = P[Y^{A=0} = 1|M = 1]$,

so that $\Delta_{CE} = 0$. This is generally false, as discussed in the first paragraph of Section 2 above, and illustrated in Figure 2.

4.3 Alternative proofs of $Y^{A=a} \not\perp\!\!\!\perp A|M$ under this model

Under this simple model, the fact that $Y^{A=a} \not\perp\!\!\!\perp A|M$ can also be shown by a simple reduction to absurdity argument. If $Y^{A=a}$ was independent of A given M, we would get the following chain of equalities:

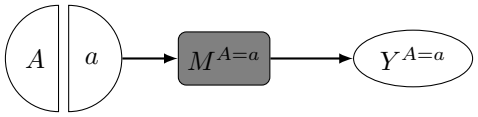
$$\begin{aligned}
P[Y^{A=a}=1] &= \sum_m P[Y^{A=a}=1|M=m]P[M=m] \quad \text{by the tower rule} \\
&= \sum_m P[Y^{A=a}|M=m, A=a]P[M=m] \quad \text{if } Y^{A=a} \perp\!\!\!\perp A|M \\
&= \sum_m P[Y=1|M=m, A=a]P[M=m] \quad \text{by consistency} \\
&= \sum_m P[Y=1|M=m]P[M=m], \quad \text{because } Y \perp\!\!\!\perp A|M \\
&= P[Y=1] \quad \text{by the tower rule.}
\end{aligned}$$

Therefore, the assumption $Y^{A=a} \perp\!\!\!\perp A|M = 1$ yields $P[Y^{A=1} = 1] - P[Y^{A=0} = 1] = P[Y = 1] - P[Y = 1] = 0$, which contradicts Equation (3) . This completes our reduction to absurdity argument and establishes that $Y^{A=a} \perp\!\!\!\perp A|M = 1$ is generally false under the model of Figure 4.

Further observe that our chain of equalities does not lead to any contradiction in the absence of either the arrow pointing from A to M or the arrow pointing from M to Y: these are actually two special cases of the model for which $Y^{A=a} \perp\!\!\!\perp A|M = 1$ does hold. This is why we observed $\Delta_{CE} = \Delta_{Sp}$ in the absence of interaction terms in the Y-model and if, in addition, $\beta_M = 0$; see the top row of Figure 2 as well as the top two rows of Figure 2 in our Letter.

The fact that $Y^{A=a} \not\perp\!\!\!\perp A|M$ can also be shown using the SWIG approach.[4] Figure 5 presents the SWIT corresponding to the intervention $do(A = a)$. From this representa-

Figure 5: The SWIT resulting from the intervention $do(A = a)$ in the simplified causal model of Figure 4.



tion, it directly follows that $A \perp\!\!\!\perp (Y^{A=a}, M^{A=a})$, and therefore that $Y^{A=a} \perp\!\!\!\perp A | M^{A=a}$; see Section 3.5.3 in[4]. Then, the following holds

$$\begin{aligned} P(Y = 1|M = 1, A = a) &= P(Y^{A=a}|M^{A=a} = 1, A = a) \quad \text{by consistency} \\ &= P(Y^{A=a}|M^{A=a} = 1) \quad \text{since } Y^{A=a} \perp\!\!\!\perp A|M^{A=a}. \quad (4) \end{aligned}$$

Observe that the random sets $\{M^{A=a} = 1\}$ and $\{M = 1\}$ are generally different. On the one hand, $\{M^{A=0} = 1\}$ consists of the individuals who would have suffered from diabetes in the counterfactual world $\Omega^{A=0}$ that we would have observed had obesity been eliminated. On the other hand, $\{M = 1\}$ consists of the individuals with diabetes in the actual world, among whom some are obese and others are not. If obesity causes diabetes it is clear that $\{M^{A=1} = 1\} \neq \{M^{A=0} = 1\}$ and then that $\{M = 1\}$ differs from $\{M^{A=a} = 1\}$, for $a \in \{0, 1\}$. To be more concrete, consider the generative model described in Section 3.1 above. Set $\beta_A = \beta_U = \alpha_U = 0$ and $\beta_{AM} = \beta_{AU} = \beta_{AUM} = \beta_{UM} = \alpha_{AU} = \alpha_{UM} = 0$ to ensure that there is no arrow pointing from A to Y and no confounder U. Then, we have $\{M^{A=0} = 1\} = \{\varepsilon_M \leq \text{expit}(\alpha_0)\}$, so that $P(M^{A=0} = 1) = \text{expit}(\alpha_0)$. Moreover, $\{M^{A=1} = 1\} = \{\varepsilon_M \leq \text{expit}(\alpha_0 + \alpha_A)\}$, so that $P(M^{A=1} = 1) = \text{expit}(\alpha_0 + \alpha_A)$. As for $\{M = 1\}$, we have

$$\begin{aligned} \{M = 1\} &= \{\varepsilon_M \leq \text{expit}(\alpha_0 + \alpha_A A)\} \\ &= (\{\varepsilon_M \leq \text{expit}(\alpha_0 + \alpha_A)\} \cap \{\varepsilon_A \leq p_A\}) \cup (\{\varepsilon_M \leq \text{expit}(\alpha_0)\} \cap \{\varepsilon_A > p_A\}), \end{aligned}$$

so that $P(M = 1) = p_A \text{expit}(\alpha_0 + \alpha_A) + (1 - p_A)\text{expit}(\alpha_0)$. Therefore, $\{M = 1\} \neq \{M^{A=0} = 1\}$, $\{M = 1\} \neq \{M^{A=1} = 1\}$ and $\{M^{A=0} = 1\} \neq \{M^{A=0} = 1\}$ as soon as $\alpha_A \neq 0$ in this simplified setting.

Because $\{M^{A=a} = 1\}$ and $\{M = 1\}$ are generally different, quantities $P[Y^{A=a} = 1|M^{A=a} = 1]$ and $P[Y^{A=a} = 1|M = 1]$ are generally different too. In view of (4), this yields $P[Y^{A=a} = 1|M = 1] \neq P(Y = 1|M = 1, A = a)$, and then $Y^{A=a} \not\perp\!\!\!\perp A|M$.

5 Discussion

Obesity is widely considered as a cause of early death. With the notations used here, this means that the causal odds-ratio,

$$[P(Y^{A=1} = 1)/P(Y^{A=1} = 0)]/[P(Y^{A=0} = 1)/P(Y^{A=0} = 0)],$$

is superior to 1. However, several observational studies reported an observed odds-ratio among individuals with diabetes or heart failure [10,11,12] less than one, suggesting that $OR_{AS} < 1$. This observation could indeed be considered as paradoxical if the observed odds-ratio among individuals with chronic disease was a consistent estimate of OR_{CE}. However, this is not the case because M is a descendent of A, as already suggested in the literature [13]. Contrary to what Sperrin et al. reported [1], we show that the difference between OR_{AS} and OR_{CE} can be sensible even under the simple causal model they considered. In addition, we show that OR_{AS}, OR_{CDE} and $OR_{CDE|A=1,M=1}$ share some similarities, but are still different. In particular, by considering additional interaction terms in the generative model proposed by Sperrin et al., we exhibited configurations where $OR_{AS} < 1$ while $OR_{CE} > 1$, $OR_{CDE} > 1$ and $OR_{CDE|A=1,M=1} > 1$. Therefore, estimates of OR_{AS} should be regarded with caution since they can not be related to any meaningful causal effects. Furthermore,

the confounder U has to be observed in order to estimate the causal quantities OR_{CDE} and $OR_{CDE|A=1,M=1}$. As for OR_{CE}, it can not be identified from the distribution of (A, M, U, Y) without further assumptions on the causal model.

We shall add that even if we could estimate OR_{CDE} and OR_{CE}, these quantities might not be appropriate to answer the question of whether weight loss would be beneficial for an obese patient with diabetes or heart failure[10]. The risk of early death for such a patient is $P(Y = 1|A = 1, M = 1) = P(Y^{A=1} = 1|A = 1, M = 1)$, by consistency. But what would be his risk after a weight loss? If it can be assumed that his risk would be the one he would have had in the counterfactual world $\Omega^{A=0}$, that is, his risk had he never been obese, then it is simply $P(Y^{A=0}|A = 1, M = 1)$. Because $P(Y^{A=0} = 1|A = 1, M = 1) = P(Y^{A=0,M=M^{A=0}} = 1|A = 1, M = 1)$, it is noteworthy that this assumption implies that this patient might be cured of diabetes after his weight loss. Under this assumption, the quantity of interest is therefore

$$\begin{aligned}
&P(Y = 1|A = 1, M = 1) - P(Y^{A=0} = 1|A = 1, M = 1) \\
&= P(Y^{A=1} = 1|A = 1, M = 1) - P(Y^{A=0} = 1|A = 1, M = 1).
\end{aligned}$$

It is related to Δ_{CE}, but also to the excess fraction and, under some assumptions, to the attributable fraction and the probability of disablement[2,5]. However, if weight loss is unlikely to cure this patient of diabetes, his risk after a weight loss might rather be $P[Y^{A=0,M=1} = 1|A = 1, M = 1]$. Then, the quantity of interest would be

$$\begin{aligned}
&P(Y = 1|A = 1, M = 1) - P[Y^{A=0,M=1} = 1|A = 1, M = 1] \\
&= P(Y^{A=1,M=1} = 1|A = 1, M = 1) - P(Y^{A=0,M=1} = 1|A = 1, M = 1) \\
&= \Delta_{CDE|A=1,M=1}.
\end{aligned}$$

Lastly, because there is no unique and well-defined intervention resulting in weight loss (or weight gain), causal inference on observational data is a particularly complicated

task when dealing with obesity[7]. As a matter of fact, to answer the question whether weight loss would be beneficial for obese patients with diabetes, a safer roadmap would be first specifying the envisaged intervention(s) which might result in weight loss, and then planning a randomized interventional study.

References

[1] M. Sperrin, J. Candlish, E. Badrick, A. Renehan, and I. Buchan, "Collider bias is only a partial explanation for the obesity paradox:," *Epidemiology*, vol. 27, no. 4, pp. 525–530, 2016.

[2] J. Pearl, *Causality: models, reasoning, and inference*. Cambridge, U.K. ; New York: Cambridge University Press, 2000.

[3] M. A. Hernan and J. M. Robins, *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming.

[4] T. S. Richardson and J. M. Robins, "Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality," *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, vol. 128, no. 30, p. 2013, 2013.

[5] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, no. 0, pp. 96–146, 2009.

[6] J. Pearl, "An introduction to causal inference," *The International Journal of Biostatistics*, vol. 6, no. 2, 2010.

[7] M. A. Hernán and S. L. Taubman, "Does obesity shorten life? The importance of well-defined interventions to answer causal questions," *International Journal of Obesity*, vol. 32, pp. S8–S14, 2008.

[8] K. Imai, L. Keele, and D. Tingley, "A general approach to causal mediation analysis.," *Psychological Methods*, vol. 15, no. 4, pp. 309–334, 2010.

[9] T. J. VanderWeele and S. Vansteelandt, "Odds Ratios for Mediation Analysis for a Dichotomous Outcome," *American Journal of Epidemiology*, vol. 172, pp. 1339–1348, Dec. 2010.

[10] S. Anker and S. Haehling, "The obesity paradox in heart failure: accepting reality and making rational decisions," *Clinical Pharmacology & Therapeutics*, vol. 90, no. 1, pp. 188–190, 2011.

[11] M. R. Carnethon, P. J. D. De Chavez, M. L. Biggs, C. E. Lewis, J. S. Pankow, A. G. Bertoni, S. H. Golden, K. Liu, K. J. Mukamal, B. Campbell-Jenkins, *et al.*, "Association of weight status with mortality in adults with incident diabetes," *JAMA*, vol. 308, no. 6, pp. 581–590, 2012.

[12] A. Oreopoulos, R. Padwal, K. Kalantar-Zadeh, G. C. Fonarow, C. M. Norris, and F. A. McAlister, "Body mass index and mortality in heart failure: a meta-analysis," *American Heart Journal*, vol. 156, no. 1, pp. 13–22, 2008.

[13] M. Lajous, A. Bijon, G. Fagherazzi, M.-C. Boutron-Ruault, B. Balkau, F. Clavel-Chapelon, and M. A. Hernán, "Body mass index, diabetes, and mortality in french women: explaining away a "paradox"," *Epidemiology (Cambridge, Mass.)*, vol. 25, no. 1, p. 10, 2014.