OPEN

# Collider Bias Is Only a Partial Explanation for the Obesity Paradox

*Matthew Sperrin,*[a] *Jane Candlish,*[a] *Ellena Badrick,*[a] *Andrew Renehan,*[b] *and Iain Buchan*[a]

**Background:** "Obesity paradox" refers to an association between obesity and reduced mortality (contrary to an expected increased mortality). A common explanation is collider stratification bias: unmeasured confounding induced by selection bias. Here, we test this supposition through a realistic generative model.

**Methods:** We quantify the collider stratification bias in a selected population using counterfactual causal analysis. We illustrate the bias for a range of scenarios, describing associations between exposure (obesity), outcome (mortality), mediator (in this example, diabetes) and an unmeasured confounder.

**Results:** Collider stratification leads to biased estimation of the causal effect of exposure on outcome. However, the bias is small relative to the causal relationships between the variables.

**Conclusions:** Collider bias can be a partial explanation of the obesity paradox, but unlikely to be the main explanation for a reverse direction of an association to a true causal relationship. Alternative explanations of the obesity paradox should be explored. See Video Abstract at http://links.lww.com/EDE/B51.

(*Epidemiology* 2016;27: 525–530)

"Obesity paradox" is the term given to the finding that, in certain populations, people who are obese seem to live longer. This has been observed in patients with coronary artery disease,[1] heart failure,[2] and type 2 diabetes,[3,4] among others.

Proposed explanations for the paradox include[5]: body fat helping patients survive periods of low nutrition; the nonobese population including patients who have lost weight as a result of more severe illness; body mass index (BMI) poorly representing body fat[6]; BMI cut-offs not being appropriate[7]; and obese people being diagnosed earlier.

This article focuses on the collider stratification bias[5,8] explanation: a correlation induced between an exposure and confounder when stratifying on a third variable (collider) that is associated with, and downstream from, both.[9] If the confounder also affects the outcome, conditioning on the collider can induce a false, strengthened, or reversed association between exposure and outcome.

Existing literature has demonstrated that collider stratification bias can occur in principle.[5,8] However, it is not known whether the conditions under which effect reversal occurs are realistic. The aim of this article is to investigate the plausibility of collider stratification bias as an explanation for the obesity paradox under a realistic generative model.

## METHODS

### Derivation

We give a general description of collider stratification bias, using the obesity paradox to illustrate, beginning with some definitions from counterfactual causal analysis.[10,11] Referring to Figure 1, our interest is in the relationship between the exposure $A$ (e.g., obesity) and the outcome $Y$ (e.g. mortality), complicated by a mediator $M$ (e.g., diabetes status) and a confounder $U$, which may be unmeasured. $U$ is assumed (unconditionally) independent of $A$. Suppose that $U$ and $A$ have distributions $F_U$ and $F_A$, respectively. While our derivations allow the variables $U$ and $A$ to take any form, the mathematics is clearer if we consider the binary case, with $P[U=1]=p_U$, $P[A=1]=p_A$, and the other two variables generated by regression equations:

$$g_M\left(E[M \mid A,U]\right)=\alpha_0+\alpha_A A+\alpha_U U+\alpha_{AU}AU \quad (1)$$

$$g_Y\left(E[Y \mid M,A,U]\right)=\beta_0+\beta_M M+\beta_U U+\beta_A A+\beta_{AM}AM, \quad (2)$$

with $g_M$, $g_Y$ = logit, with inverse: $\text{expit}(x)=\left(1+\exp(-x)\right)^{-1}$.

We are interested in the causal effect of $A$ on $Y$ (obesity on mortality), conditioned on $M$ being at level $m$ (diabetes status), i.e., a comparison in which $A$ is set (counterfactually) to level $a$ or $a^*$ (e.g., obese or nonobese):

$$\Delta_{\text{CE}}=\delta\left(E\left[Y^{A=a}|M=m\right],\, E\left[Y^{A=a^*}|M=m\right]\right)$$

Submitted 8 June 2015; accepted 31 March 2016.

From the [a]Health eResearch Centre, Farr Institute, and [b]Institute of Cancer Sciences, Manchester Academic Health Science Centre, University of Manchester, Manchester, United Kingdom.

Supported by the University of Manchester's Health eResearch Centre (HeRC) funded by the Medical Research Council Grant MR/K006665/1.

The authors report no conflicts of interest.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Matthew Sperrin, Health eResearch Centre, Farr Institute, University of Manchester, Manchester M13 9PL, United Kingdom. E-mail: matthew.sperrin@manchester.ac.uk.



ISSN: 1044-3983/16/2704-0525
DOI: 10.1097/EDE.0000000000000493

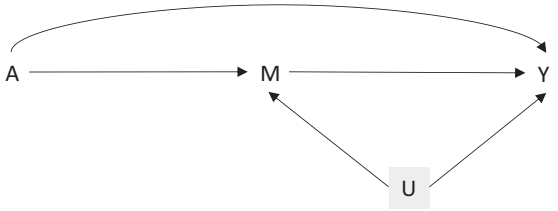


**FIGURE 1.** Illustration of collider stratification bias.

where $\delta$ represents a difference between the two expectations. However, we calculate the association:

$$\Delta_{\mathrm{AS}} = \delta\big(E[Y|M=m, A=a], E[Y|M=m, A=a^*]\big),$$

in which we compare individuals *observed* at exposure levels $a$ and $a^*$. Effect sizes of interest may be a (log) risk ratio or (log) odds ratio. The obesity paradox explained by collider stratification bias argument follows from the noninequality of the above measures. A scenario of particular interest is when the association is the reverse of the causal effect.

The nonequality of $\Delta_{\mathrm{CE}}$ and $\Delta_{\mathrm{AS}}$ is possible because

$$E\big[Y^{A=a}|M=m\big] = \sum_u E\big[Y^{A=a}|m,u\big]P[u|m] = \sum_u E[Y|m,a,u]P[u|m],$$

while

$$E[Y|M=m, A=a] = \sum_u E[Y|m,a,u]P[u|m,a].$$

In general, $P[u|m] \neq P[u \mid m,a]$ because conditioning on $M$ induces a relationship between $U$ and $A$.

Following similar lines to the study of Vanderweele,[12] the above can be computed:

$$E[Y|m,a,u] = g_Y^{-1}\big(\beta_0 + \beta_M m + \beta_U u + \beta_A a + \beta_{AM} am\big),$$

$$P[u|m,a] = \frac{P[m|u,a]P[u \mid a]}{P[m \mid a]} = \frac{g_M^{-1}\big(\alpha_0 + \alpha_A a + \alpha_U u + \alpha_{AU} au\big)P[u]}{\sum_u g_M^{-1}\big(\alpha_0 + \alpha_A a + \alpha_U u + \alpha_{AU} au\big)P[u]},$$

$$P[u \mid m] = \frac{P[m|u]P[u]}{P[m]} = P[u]\frac{\sum_\alpha P[m|u,\alpha]P[a|u]}{\sum_{a,u} P[m|u,a]P[a|u]P[u]} = P[u]\frac{\sum_\alpha g_M^{-1}\big(\alpha_0 + \alpha_A a + \alpha_U u + \alpha_{AU} au\big)P[a]}{\sum_{a,u} g_M^{-1}\big(\alpha_0 + \alpha_A \alpha + \alpha_U u + \alpha_{AU} au\big)P[a]P[u]}.$$

With a logit link, for $\Delta_{\mathrm{CE}}$ we have

$$P\big[Y^{A=1}=1|M=1\big] = \mathrm{expit}\big(\beta_0 + \beta_M + \beta_U + \beta_A + \beta_{AM}\big)\, p_{U|M} + \mathrm{expit}\big(\beta_0 + \beta_M + \beta_A + \beta_{AM}\big)(1-p_{U|M}),$$

and

$$P\big[Y^{A=0}=1|M=1\big] = \mathrm{expit}\big(\beta_0 + \beta_M + \beta_U\big)\, p_{U|M} + \mathrm{expit}\big(\beta_0 + \beta_M\big)(1-p_{U|M}),$$

where

$$p_{U|M} = P[U=1|M=1] = \frac{p_U\{\mathrm{expit}\big(\alpha_0 + \alpha_A + \alpha_U + \alpha_{AU}\big)\, p_A + \mathrm{expit}\big(\alpha_0 + \alpha_U\big)\big(1-p_A\big)\}}{p_U\big\{\mathrm{expit}\big(\alpha_0 + \alpha_A + \alpha_U + \alpha_{AU}\big)p_A + \mathrm{expit}\big(\alpha_0 + \alpha_U\big)\big(1-p_A\big)\big\} + (1-p_U)\{\mathrm{expit}\big(\alpha_0 + \alpha_A\big)p_A + \mathrm{expit}\big(\alpha_0\big)\big(1-p_A\big)\}}.$$

In particular $P\big[Y^{A=1}=1|M=1\big] = P\big[Y^{A=0}=1|M=1\big]$ if $\beta_A = \beta_{AM} = 0$.

For $\Delta_{\mathrm{AS}}$:

$$P[Y=1|M=1,A=1] = \left\{\begin{array}{l}\mathrm{expit}\big(\beta_0 + \beta_M + \beta_U + \beta_A + \beta_{AM}\big)\mathrm{expit}\big(\alpha_0 + \alpha_A + \alpha_U + \alpha_{AU}\big) \\ p_U + \mathrm{expit}(\beta_0 + \beta_M + \beta_A + \beta_{AM})\mathrm{expit}\big(\alpha_0 + \alpha_A\big)(1-p_U) \\ \hline \mathrm{expit}\big(\alpha_0 + \alpha_A + \alpha_U + \alpha_{AU}\big) \\ p_U + \mathrm{expit}\big(\alpha_0 + \alpha_A\big)(1-p_U)\end{array}\right\};$$

similarly,

$$P[Y=1|M=1,A=0] = \left\{\begin{array}{l}\mathrm{expit}\big(\beta_0 + \beta_M + \beta_U\big)\mathrm{expit}\big(\alpha_0 + \alpha_U\big) \\ p_U + \mathrm{expit}\big(\beta_0 + \beta_M\big)\mathrm{expit}\big(\alpha_0\big)(1-p_U) \\ \hline \mathrm{expit}\big(\alpha_0 + \alpha_U\big)p_U + \mathrm{expit}\big(\alpha_0\big)(1-p_U)\end{array}\right\}.$$

Both the causal effect and association are a weighted average of $\mathrm{expit}\big(\beta_0 + \beta_M + \beta_U\big)$ and $\mathrm{expit}\big(\beta_0 + \beta_M\big)$, but the weights differ depending on the status of $A$ and $U$, so a spurious association may be observed.

## Illustration

The model is described by Figure 1, and regression Equations (1, 2). We supposed that the only data available are those with $M=1$ (e.g., those with diabetes). We visualized the discrepancy between the association and causal effect for a range of parameter values. The collider bias variables $\alpha_A, \alpha_U, \alpha_{AU}$, and $\beta_U$ were varied over a grid from −3 to 3; this range captures the salient features, and covers the range of parameters that may reasonably be observed in practice. We considered two scenarios for $\beta_A$ : no causal effect ($\beta_A = 0$), and some causal effect ($\beta_A = 1$). Throughout we set $p_U = p_A = 0.5$, $\alpha_0 = -\frac{1}{2}\left(\alpha_A + \alpha_U + \frac{1}{2}\alpha_{AU}\right)$, and $\beta_0 = -\frac{1}{2}(\beta_U + \beta_A)$, so that the prevalences of all variables remain close to 50% (e.g., half the

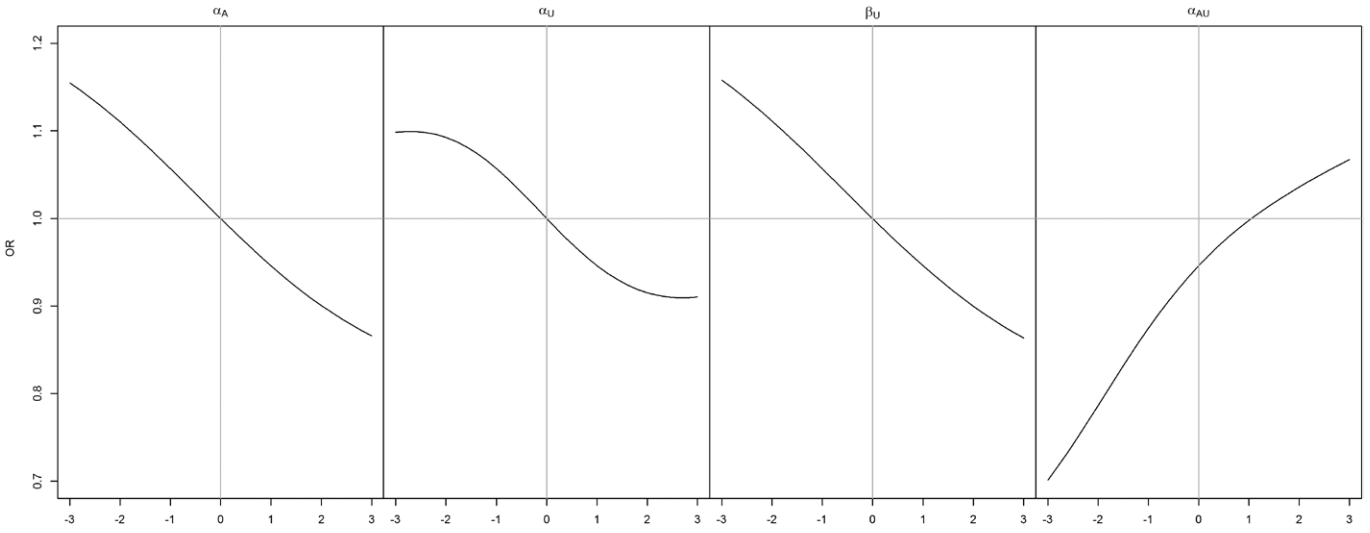


**FIGURE 2.** Association (OR) between *A* and *Y* in the null case for varying values of $\alpha_A, \alpha_U, \beta_U$, and $\alpha_{AU}$. In each panel, along the *x* axis, one of these variables is varied from −3 to 3 (*left panel*: $\alpha_A$, *mid-left panel*: $\alpha_U$, *mid-right panel*: $\beta_U$, *right panel*: $\alpha_{AU}$), and the other parameters are set to default values.

population are obese). We set $\beta_M = \beta_{AM} = 0$, without loss of generality.

We presented all effect sizes as odds ratios. The illustrations were produced using R 3.1.0[13]; code is in eAppendix 1 (http://links.lww.com/EDE/B48).

## RESULTS

Figures 2 and 3 visualize the association between *A* and *Y* when there is no causal effect. Figure 2 looks at the impact of each of the collider stratification bias variables $\alpha_A, \alpha_U, \alpha_{AU}$, and $\beta_U$, one at a time; in each case the remaining variables are set to 1, except $\alpha_{AU} = 0$. The left panel of Figure 2 shows that when $\alpha_A$ is positive (with $\alpha_U = 1, \beta_U = 1, \alpha_{AU} = 0$), the observed association between *A* and *Y* is negative. This represents a bias because the causal effect is zero. In the diabetes example, this means that in a diabetes population where *A* (obesity) and *U* both increase the risk of *M* (diabetes), and *U* also increases the risk of *Y* (death), but *A* has no effect on *Y* except through *M*, we observe a negative association between *A* and *Y*. Similar results are seen for the other parameters. In Figure 3, each row in the lattice corresponds to a value of $\alpha_A$, while each column corresponds to a value of $\beta_U$. Within each graph, $\alpha_U$ is varied from −3 to 3, and we consider no interaction ($\alpha_{AU} = 0$, solid line) and antagonistic interaction ($\alpha_{AU} = -1$, dotted line). The bottom right panels of Figure 3 illustrates that when $\alpha_A, \alpha_U$, and $\beta_U$ parameters are positive, the association between *A* and *Y* becomes negative.

Figures 4 and 5 visualize the association, and causal effect, between *A* and *Y* when there is a causal effect, $\beta_A = 1$. Obesity paradox occurs when the association has the opposite sign to the causal effect. The direction of bias between the association and causal effect are the same as when the causal effect was zero. Notably, obesity paradox is happening only for configurations such as $\alpha_A = 3$, $\alpha_U = 3$, $\beta_U = 3$ (see the bottom right panel of Figure 5), i.e., when all the parameters on the confounding pathway are substantially larger than the causal effect. The association in the reverse direction is small, amplified slightly by antagonism between *A* and *U* ($\alpha_{AU} = -1$).

## DISCUSSION

Contrary to much recent literature, our results suggest that collider bias alone cannot fully explain the obesity paradox, with only small discrepancies between the association and the causal effect observed. For large discrepancies to occur (e.g., for the association to reverse the causal effect), the parameters on the collider bias pathway must be large compared with the true causal effect. This could only happen if the true causal effect is small, and therefore unlikely to be important; or the effect of the unmeasured confounder on both the mediator and the outcome is very large, therefore unlikely to be missed from the analysis.

Glymour and Vittinghoff[14] also demonstrated that collider bias must be very strong to lead to an association that reverses the causal effect, and Greenland[15] gave a formula for calculating the maximum observable bias. Banack and Kaufman[16] studied the strength of collider bias required to reverse a particular causal effect. While they concluded that such a reversal was plausible, strong relationships along the collider stratification bias pathway are nevertheless required. Collider stratification bias does not apply when the population is unselected, so our finding is supported by a similar protective effect of obesity in the general population.[17]

For certain nonzero configurations of the model parameters, there is no bias in the estimation of the causal effect (e.g., the crossing of the *x* axis in right panel, Figure 2). This is unfaithfulness, which occurs when a multiplicative model is induced in the risk scale.[14,18,19]

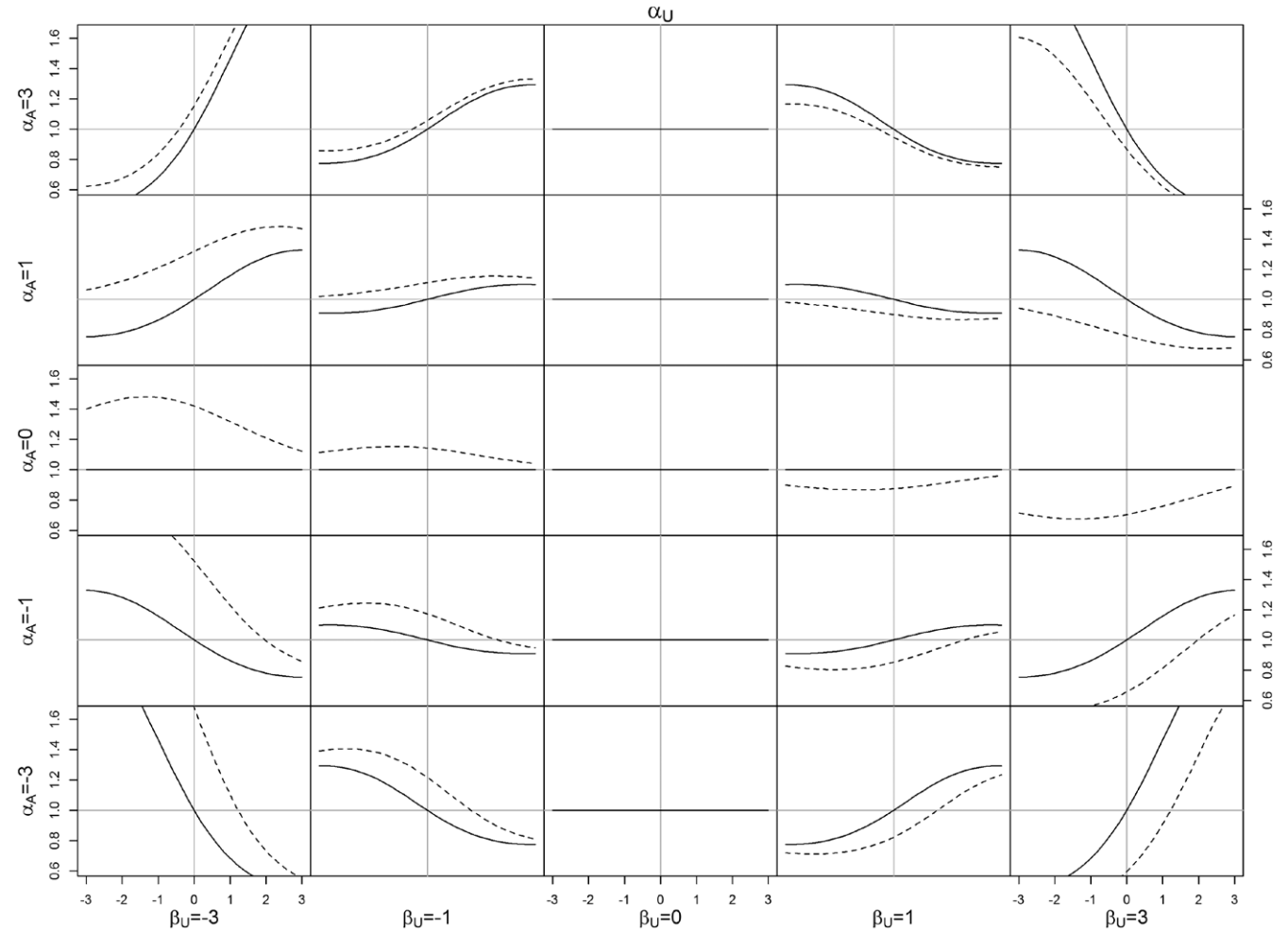


**FIGURE 3.** Association (OR) between *A* and *Y* in the null case for varying values of $\alpha_A$,$\alpha_U$, and $\beta_U$, without interaction ($\alpha_{AU} = 0$, *solid line*) and with interaction ($\alpha_{AU} = 1$, *dotted line*). Each column (*row*) in the lattice corresponds to the given value of $\beta_U$ ($\alpha_A$). Within each subgraph, along the *x* axis, $\alpha_U$ is varied from −3 to 3.

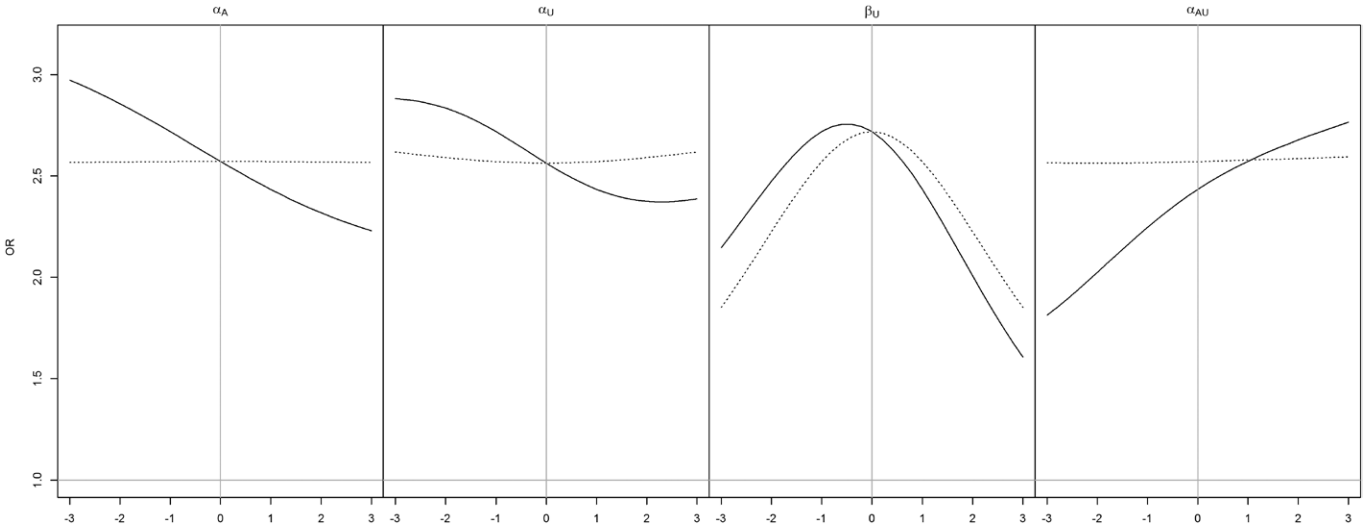


**FIGURE 4.** Association (OR) between *A* and *Y* (*solid line*) versus causal effect (log odds) of *A* on *Y* (*dashed line*) for a range of values of $\alpha_A$,$\alpha_U$, $\beta_U$, and $\alpha_{AU}$. In each panel, along the *x* axis, one of these variables is varied from −3 to 3 (*left panel*: $\alpha_A$, *mid-left panel*: $\alpha_U$, *mid-right panel*: $\beta_U$, *right panel*: $\alpha_{AU}$), and the other parameters are set to default values.

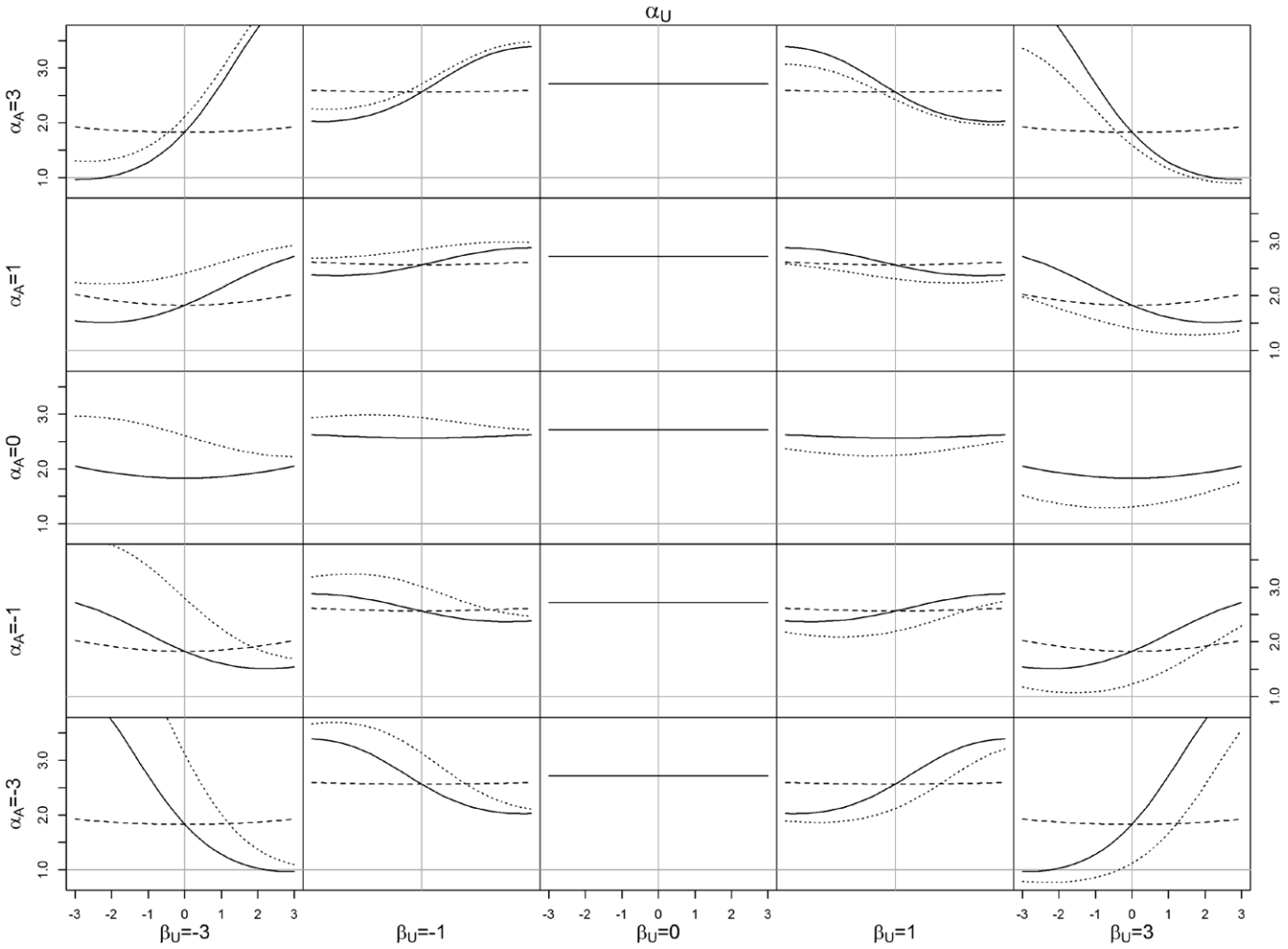


**FIGURE 5.** Association (OR) between *A* and *Y* versus causal effect (log odds) of *A* on *Y* for varying values of $\alpha_A, \alpha_U$, and $\beta_U$, with $\beta_B = 1$. Each column (*row*) in the lattice corresponds to the given of $\beta_U$ ($\alpha_A$). Within each subgraph, along the *x* axis, $\alpha_U$ is varied from −3 to 3. Association without interaction ($\alpha_{AU} = 0$): *solid line*; association with interaction ($\alpha_{AU} = 1$): *dotted line*; causal effect: *dashed line*.

A strength of our study is that our findings are based on mathematical results rather than simulations. However, we restricted to binary variables. Further study is needed to extend this: one context of interest within the obesity paradox is time to event outcome (death), and continuous exposure (BMI).

We have given a simple exposition here, based on a minimal set of four variables. Two of these variables (*U* and *A*) were assumed independent; however, dependence between these variables may affect the degree of the collider bias.[16] There may be multiple confounding variables; we have considered only one. There may be latent subtypes of the mediating disease.[20] Finally, we have not considered the time-varying nature of obesity.

When examining the relationship between an exposure and outcome in a subpopulation, the real interest is in whether this relationship differs from the population as a whole, i.e., the relationship is moderated by the mediator. This can only be assessed by modeling the whole population, with interaction terms between exposure and moderator. However, statistical interaction does not necessarily imply a true biological interaction.[21]

Our results show that the paradoxical observation of a protective effect of obesity on mortality is unlikely to be fully explained by collider stratification bias.

## REFERENCES

1. Romero-Corral A, Montori VM, Somers VK, et al. Association of bodyweight with total mortality and with cardiovascular events in coronary artery disease: a systematic review of cohort studies. *Lancet*. 2006;368:666–678.
2. Oreopoulos A, Padwal R, Kalantar-Zadeh K, Fonarow GC, Norris CM, McAlister FA. Body mass index and mortality in heart failure: a meta-analysis. *Am Heart J*. 2008;156:13–22.
3. Carnethon MR, De Chavez PJ, Biggs ML, et al. Association of weight status with mortality in adults with incident diabetes. *JAMA*. 2012;308:581–590.
4. Zhao W, Katzmarzyk PT, Horswell R, et al. Body mass index and the risk of all-cause mortality among patients with type 2 diabetes. *Circulation*. 2014; 130: 2143–2151.

5. Banack HR, Kaufman JS. The obesity paradox: understanding the effect of obesity on mortality among individuals with cardiovascular disease. *Prev Med*. 2014;62:96–102.
6. Rothman KJ. BMI-related errors in the measurement of obesity. *Int J Obes (Lond)*. 2008;32(suppl 3):S56–S59.
7. Dixon JB, Egger GJ, Finkelstein EA, Kral JG, Lambert GW. 'Obesity paradox' misunderstands the biology of optimal weight throughout the life cycle. *Int J Obes (Lond)*. 2015;39:82–84.
8. Preston SH, Stokes A. Obesity paradox: conditioning on disease enhances biases in estimating the mortality risks of obesity. *Epidemiology*. 2014;25:454–461.
9. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol*. 2010;39:417–420.
10. Pearl J. *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge University Press; 2009. London, UK.
11. Hernan MA, Robins JM. *Causal Inference*. Boca Raton, FL: Chapman & Hall/CRC, forthcoming; 2016.
12. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*. 2010;21:540–551.
13. R Core Team. R: A Language and Environment for Statistical Computing. 2014.
14. Glymour MM, Vittinghoff E. Commentary: selection bias as an explanation for the obesity paradox: just because it's possible doesn't mean it's plausible. *Epidemiology*. 2014;25:4–6.
15. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003;14:300–306.
16. Banack HR, Kaufman JS. Does selection bias explain the obesity paradox among individuals with cardiovascular disease? *Ann Epidemiol*. 2015;25:342–349.
17. Flegal KM, Kit BK, Orpana H, Graubard BI. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. *JAMA* 2013;309:71–82.
18. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10:37–48.
19. Greenland S, Mansournia MA. Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. *Eur J Epidemiol*. 2015;30:1101–1110.
20. Lajous M, Bijon A, Fagherazzi G, et al. Body mass index, diabetes, and mortality in French women: explaining away a "paradox". *Epidemiology*. 2014;25:10–14.
21. Berzuini C, Dawid AP. Deep determinism and the assessment of mechanistic interaction. *Biostatistics*. 2013;14:502–513.