International Journal of Obesity (2008) 32, S8–S14
 0307-0565/08 $30.00
www.nature.com/ijo

ORIGINAL ARTICLE

Does obesity shorten life? The importance of well-defined interventions to answer causal questions

MA Hernán[1,2] and SL Taubman[1,3]

[1]Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA; [2]Harvard-MIT Division of Health Sciences and Technology, Boston, MA, USA and [3]Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

Many observational studies have estimated a strong effect of obesity on mortality. In this paper, we explicitly define the causal question that is asked by these studies and discuss the problems associated with it. We argue that observational studies of obesity and mortality violate the condition of consistency of counterfactual (potential) outcomes, a necessary condition for meaningful causal inference, because (1) they do not explicitly specify the interventions on body mass index (BMI) that are being compared and (2) different methods to modify BMI may lead to different counterfactual mortality outcomes, even if they lead to the same BMI value in a given person. Besides precluding the estimation of unambiguous causal effects, this violation of consistency affects the ability to address two additional conditions that are also necessary for causal inference: exchangeability and positivity. We conclude that consistency violations not only preclude the estimation of well-defined causal effects but also compromise our ability to estimate ill-defined causal effects.
International Journal of Obesity (2008) **32,** S8–S14; doi:10.1038/ijo.2008.82

Keywords: causal effects; causal inference; interventions; confounding; body mass index

Introduction

Several observational studies[1–4] have estimated that over 100 000 excess deaths are attributable to obesity (defined as body mass index (BMI) $\geqslant 30$) and overweight ($25 \leqslant$ BMI < 30) each year in the United States. These estimates place high BMI as the second preventable cause of death, after cigarette smoking, and thus as one of the most important public health problems of the country.[2] In this paper, we explicitly define the causal question that is asked by these mortality studies and discuss the problems associated with it.

We start by telling a tale of two policy makers—one prefers randomized experiments, the other prefers observational studies—leading to the apparently counterintuitive conclusion that observational studies are better suited than randomized experiments to answer causal questions regarding the health effects of obesity. As explained in the next section, the source of this paradox is a violation of the condition of consistency of counterfactuals in the observational study. The next two sections explore the effects of lack of consistency in observational studies of obesity on two additional conditions for causal inference: exchangeability and positivity. We conclude that consistency violations not only preclude the estimation of well-defined causal effects but also compromise our ability to estimate even ill-defined causal effects. Although recent research has drawn attention to the potentially different associations between obesity (BMI $\geqslant 30$) and overweight ($25 \leqslant$ BMI < 30) with mortality, for simplicity we will talk about obesity, rather than overweight and obesity, in our hypothetical examples.

Correspondence: Dr MA Hernán, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA.
E-mail: miguel_hernan@post.harvard.edu

A tale of two policy makers

You were hired by an enlightened, but despotic, king to estimate the effect of obesity on mortality in his country. Aware of the limitations of observational studies, the king made it clear that only evidence from randomized experiments was acceptable as a basis for his policy making. Your work was to be carried out free of temporal, financial, logistical and ethical constraints. You decided to conduct a large randomized trial in which one million subjects (of the king) were assigned to one of two groups. The intervention group was forced to start an intense exercise program of 1 h of strenuous physical activity per day. Compliance with the assigned regimen was strictly enforced by the police for 30 years. Those in the no-intervention group were strongly encouraged to keep their usual level of physical activity. Both

the BMI distribution and the mortality rate during the study period were lower in the intervention group. On the basis of your study results, you estimated that 100 000 annual deaths could be prevented in the kingdom by implementing your physical activity regimen.

Unknown to you, one of the coauthors of this article, Miguel, was also hired to conduct another one-million-subject randomized trial to estimate the effect of obesity on mortality over the same period in the same country. In his study, the intervention group was forced to start a comprehensive dietary intervention that limited their intake of calories and carbohydrates, whereas those in the no-intervention group were strongly encouraged to keep their normal intake of calories and carbohydrates. Both the BMI distribution and the mortality rate during the study period were lower in the intervention group. On the basis of his study results, Miguel estimated that 50 000 annual deaths could be prevented by implementing his dietary plan.

Unknown to both you and Miguel, the other coauthor of this article, Sarah, conducted yet another one-million-subject randomized trial. Her intervention group was forced to start both an exercise program (less intense than yours) and a dietary intervention (less comprehensive than Miguel's), whereas those in her no-intervention group were strongly encouraged to keep their usual physical activity and diet. Both the BMI distribution and the mortality rate during the study period were lower in the intervention group. On the basis of her study results, Sarah estimated that 120 000 annual deaths could be prevented by implementing her joint exercise and diet program.

Interestingly, even though the mortality rate in the intervention group was highest in Miguel's study and lowest in Sarah's study, all three intervention groups had the same distribution of BMI during the intervention. The distribution of BMI and the mortality rate were identical in all three no-intervention groups.

Upon completion of the randomized trials, the three of us were summoned before the king of the country.

'So,' he asked us from his throne, 'tell me. How many excess deaths are attributable to obesity?'

'It is hard to tell based on our data. The three studies had the same BMI distribution during the entire follow-up, but the mortality rates were different. There must be something about the way in which each of us decided to intervene on body weight.'

'What do you mean? I gave you unlimited resources. Can't you answer such a simple question? I need to make public health policy. What should I do?'

'Oh, that's a different question. We can help with that one. Our data indicate that a combined intervention on physical activity and diet is your best bet to reduce mortality.'

The king announces a nationwide program to encourage his subjects to be more physically active and modify their diet.

After learning about the king's scientific inclinations, the benevolent president of a neighboring state also wants to estimate the effect of obesity in his country. He, however, cannot afford to await results for 30 years (for the next election is only a couple of years away), to fund huge randomized trials such as the ones we conducted (for he runs on a low taxes ticket), or to force his citizens to comply with draconian lifestyle interventions (for judges would surely interfere). But the president has a brilliant idea, taking advantage of the country's superb, prospectively recorded, computerized health data. For all his citizens over the last 30 years, he uploads the following information to the government's Web site: (1) annual BMI measurements, (2) date of death, and (3) detailed information on lifestyle (for example, smoking, diet, physical activity) and risk factors for mortality (for example, hypertension, diabetes). Then he offers a reward to the first person who correctly estimates the excess number of deaths attributable to obesity in his country. A week later he receives a letter from a data analyst from a prestigious university with the following message: 150 000 annual excess deaths are attributable to obesity. All experts agree that the analyst used the best available statistical methods.

The president is exultant. He has found a way to avoid the huge temporal, financial, logistical and ethical difficulties associated with randomized experiments (even though the data analyst was paid more than the combined sum that the three of us received from the king). The president announces a nationwide program to encourage his people to lose weight.

Interventions and causal inference

We often hear that randomized experiments are the ideal way to answer causal questions. However, our tale describes a situation in which no randomized experiment could directly answer the causal question—What is the effect of obesity on death?—whereas an observational study seemed to be able to do so. Let us explore this apparent paradox.

Randomized trials compare the distribution of the outcome Y of interest under two or more interventions (for example, treatment regimens, plans, programs) that are randomly assigned. For example, your trial assigned subjects to either an intense physical activity program $A=1$ or to their normal exercise levels $A=0$ (in some trials, the reference intervention $A=0$ is 'no intervention'). If the distribution of the outcome Y varies by levels of the assigned intervention A, we say that A has a causal effect on Y. We can then use the magnitude of the effect of A on Y from the randomized trial to estimate the impact that such intervention would have on the population as long as the population characteristics, assignment conditions and adherence to the intervention are comparable with those in the trial. For example, in your physical activity trial, you were able to estimate that 100 000 annual deaths would be prevented by intervening on physical activity in a certain way (for example, 1 h of strenuous physical activity per day). An equivalent way to express this result is that 100 000 annual excess deaths are *attributable to not intervening on physical*

activity in a certain way. A close but not completely equivalent way to express this result is that 100 000 annual deaths are *attributable to inadequate levels of physical activity*, because the particular number of deaths (100 000 in this example) depends on the actual intervention that you studied. For example, if your intervention had been 6 h, rather than 1 h, of strenuous physical activity per day, you might have estimated that physical activity increases, rather than decreases, mortality in the general population.

You could not conclude from your trial that 100 000 annual excess deaths are attributable to obesity. Randomized trials allow us to make causal inferences about the particular interventions that they compare, but obesity was not one of the interventions compared. In fact, BMI is not an intervention at all, but rather, the potential result of many different types of interventions. In the above randomized trials, each of us chose a different method to intervene on BMI (physical activity, diet or a combination of both). In all three studies, the achieved distribution of BMI was the same, yet we estimated different effects on mortality. In fact, different effects on mortality could have logically occurred even if each of the interventions had moved every single subject's BMI to the same value, because each of the methods to change BMI may have direct effects on mortality that are not mediated through body weight. None of us could design a randomized trial to directly address the question—How many deaths are attributable to obesity?—even in the absence of temporal, financial, logistical and ethical constraints, because the answer to that question depends on the particular intervention that is implemented to eliminate obesity.

Now consider the observational study in the neighboring country. The data analyst estimated the causal effect of obesity on mortality by comparing the mortality rate in subjects who happened to be obese (BMI$\geqslant$30) with that in subjects who happened to be of normal weight (BMI in the range 18.5–25). Let us assume for now that the data analyst adjusted for all the relevant confounders. (See Consistency and Exchangeability section for a more detailed discussion.) The statement—A total of 150 000 annual deaths are attributable to obesity—is equivalent to the statement—If all obese subjects had been made to have normal weight, then 150 000 annual deaths would have been prevented—or to the statement—If we *intervene* to decrease the weight of all obese subjects to normal levels, we will prevent 150 000 annual deaths. Thus, some sort of intervention—perhaps a joint intervention on several determinants of body weight or a mixture of different interventions that vary across subjects—is implicit in observational studies that estimate the causal effect of BMI. A careful definition of this implicit intervention is crucial to give a meaning to the expression 'the causal effect of obesity.' By presenting a single observational estimate for the effect of obesity, the data analyst is either (1) assuming that all interventions on BMI have the same effect on mortality or (2) implicitly considering a complex intervention that changes the determinants of BMI in proportion to the distribution of those determinants in those who are not obese in the population. Unfortunately, both of these may lead to problems. The assumption (that all interventions on BMI have the same effects on mortality) is unlikely to hold true. The implicit intervention (on all the determinants of BMI) is likely to be too difficult to implement to be meaningful to policy makers.

Still, one could argue that there is some value in learning that 150 000 deaths could have been prevented if all obese people had been forced to be of normal weight, even if the intervention required for achieving that transformation is not well defined. This is an appealing, but risky, argument. To explain, let us be more formal.

Consistency as a fundamental condition for causal inference

Causal inference from observational data requires three key conditions: consistency, exchangeability and positivity (formally defined in the appendix). For a basic review of the assumptions of exchangeability and positivity (also known as the assumptions of no unmeasured confounding and of experimental treatment, respectively), see the introductory articles by Hernán[5] and Hernán and Robins.[6] This commentary has so far focused on the often neglected condition of consistency, which can be thought of as the condition that the causal contrast involves two or more well-defined interventions.[7,8] We will discuss how violations of consistency affect the plausibility of exchangeability and positivity. First, we need to introduce counterfactual or potential outcomes to provide a more formal definition of consistency.

For a subject participating in your randomized experiment, we define the following four variables: $A=1$ if he was assigned to the physical activity intervention (0 otherwise), $Y=1$ if he died by the end of follow-up (0 otherwise), $Y_{a=1}=1$ if he would have died had he been assigned to the intervention group $a=1$ (0 otherwise) and $Y_{a=0}=1$ if he would have died had he been assigned to the no-intervention group $a=0$ (0 otherwise). Uppercase A represents a random variable, and lowercase a represents a particular value, or realization, of that random variable. We say that the exposure A has a causal effect on the subject's outcome Y when $Y_{a=1} \neq Y_{a=0}$.

The variables $Y_{a=1}$ and $Y_{a=0}$ are referred to as counterfactual outcomes because they describe a situation that did not actually occur for some of the subjects. They are also known as potential outcomes because, for a given subject, each has a potential to be observed until treatment assignment is made. For example, if a subject was assigned to the intervention group, $A=1$, and did not die by the end of follow-up, $Y=0$, then we do not know what his outcome $Y_{a=0}$ would have been if he had been assigned to the no-intervention group. In contrast, we do know what his outcome $Y_{a=1}$ is, with his having been assigned to the intervention group. For this subject, the value of $Y_{a=1}$ is 0, the value of outcome Y that we observed. Indeed, for all

subjects with $A=1$, we know that $Y_{a=1}=Y$; this is what we mean by consistency.

Consistency may seem so obvious as to hardly deserve any attention. As a consequence, the condition of consistency is often taken for granted, and investigators tend to focus on the two other conditions for causal inference: exchangeability and positivity. Indeed, consistency is a trivial condition in randomized experiments. For example, consider a subject who was assigned to the intervention group $A=1$ in your randomized trial. By definition, it is true that, had he been assigned to the intervention, his counterfactual outcome would have been equal to his observed outcome. But the condition is not so obvious in observational studies.

Suppose that, in the observational study in the neighboring country, the data analyst compared the mortality of subjects who happened to have a BMI of 30 ($A=1$) and a BMI of 20 ($A=0$) at baseline. Now consider a study subject who had a BMI of 20 at baseline. It is not obvious that, had he been assigned to a BMI of 20 some time before baseline, his counterfactual outcome at the end of the study would have been necessarily equal to his observed outcome because there are many possible methods to assign someone to a BMI of 20. In fact, it is easy to imagine multiple ways in which a single person could reach a BMI of 20, and similarly easy to imagine that person having different mortality outcomes in each of those scenarios, despite having a BMI of 20 in each. Some of those procedures (for example, chopping off an arm, starvation, smoking) can be ruled out from our discussion because they clearly do not correspond to any interesting intervention from either a scientific or public health standpoint. We usually consider procedures (for example, diet, exercise) that correspond to interesting interventions. It is unclear whether yet other procedures (for example, gastric surgery, liposuction, genetic modification) may lead to interesting interventions.

In any case, in an observational study, we do not know the actual procedure by which each subject achieved a BMI of 20; thus, the counterfactual outcome $Y_{a=1}$ when assigned to a BMI of 20 is too vague a concept. An immediate consequence of a vague counterfactual outcome is that any causal contrast involving that counterfactual outcome will be ill defined. In other words, the data analyst can find that the mortality in subjects with a BMI of 20 differs from that in subjects with a BMI of 30, but that observed difference cannot be translated into a well-defined causal effect.

We now return to the question at the end of the last section: what is so wrong with not knowing which particular causal effect is estimated in an observational study? As long as we learn that obesity affects mortality, it may not be that important to carefully define the implicit interventions to recommend public health measures aimed at keeping BMI under 25. The problem with this argument is that lack of consistency makes it difficult to learn whether obesity really affects mortality. This is so because lack of consistency hamstrings our ability to address exchangeability and positivity.

Consistency and exchangeability

In large randomized experiments with assigned treatment A and outcome Y, measures of association between A and Y can be interpreted as measures of the effect of A on Y: association is causation. This is so because randomization ensures that the exposed ($A=1$) and the unexposed ($A=0$) groups are exchangeable: the distribution of the outcome in both groups would have been the same had the same treatment level been applied to all of them.

Exchangeability is expected in large randomized experiments, but it is unlikely to hold in many observational studies. For example, in your randomized trial of physical activity, both groups $A=1$ and $A=0$ were exchangeable because all subjects, regardless of their underlying health or lifestyle, had the same probability of being assigned to the physical activity intervention $A=1$. In contrast, in an observational study of physical activity, healthier and more health-conscious subjects may be more likely to be in the group $A=1$ (high physical activity) than in the group $A=0$. Because health status and lifestyle are common causes of physical activity and mortality, we would say that there is confounding for the effect of physical activity (that is, lack of unconditional exchangeability) and that confounding adjustment is needed. Hence, in observational studies, investigators must use their expert subject matter knowledge to identify the common causes of exposure and outcome, and measure them or their proxies (that is, the confounders L). Note that the confounders may differ with the exposure of choice, even when studying the same outcome. For example, a history of chronic obstructive pulmonary disease is a common cause of both low physical activity and mortality, but perhaps not of diet and mortality.

The hope of measuring confounders is to achieve conditional exchangeability within levels of the measured variables. The conditional exchangeability (or no unmeasured confounding) assumption allows one to estimate causal effects from observed associations.[6,9] Stratified randomization enforces conditional exchangeability, but it is impossible to check whether conditional exchangeability is met in an observational study. Investigators cannot possibly observe the subjects' counterfactual outcomes and therefore cannot be certain that their efforts to measure all confounders, which were grounded on their subject-matter knowledge, have resulted in approximate conditional exchangeability. This uncertainty is a fundamental shortcoming of causal inference from observational data.

This fundamental uncertainty is exacerbated by violations of the consistency condition. For example, consider the observational study aimed at estimating the effect of obesity on mortality. Suppose we are satisfied with learning that there is some sort of causal effect of obesity but give up on precisely characterizing it because the implicit interventions are unknowable. We still need to identify and measure the confounders for the effect of obesity on mortality so that we can achieve approximate conditional exchangeability. But

what are the common causes of obesity and mortality? Almost certainly diet and physical activity play a role—and likely have direct effects on mortality that are not mediated through BMI—and probably also asymptomatic clinical diseases and complex genetic factors (or interactions of genetic and environmental factors) currently unknown to science. Thus, when trying to estimate the effect of an ill-defined intervention by contrasting the outcome distribution between two groups of subjects who happen to differ with respect to some physiological measure (for example, BMI, low-density lipoprotein-cholesterol, CD4 cell count, C-reactive protein), it will be very hard to achieve conditional exchangeability. For example, in an obesity study, we would need to identify and measure all genetic factors that affect both body weight and mortality. This requirement imposes a heavy burden on the investigators.

Of course, conditional exchangeability cannot be guaranteed even when contrasting the outcome distribution under two well-defined interventions, as discussed above. But for many well-defined interventions (for example, drug therapy, cigarette smoking and perhaps physical activity and diet), we can reasonably argue that the major common causes of exposure and outcome do not include complex physiological or genetic processes, or if they do, their proxies (for example, indications for drug therapy) are sufficient to adjust for confounding. Confounders for well-defined interventions may then be easier to identify than those for ill-defined interventions.

In the absence of conditional exchangeability, any association between BMI and mortality may reflect not only the causal effect of BMI on mortality, but also the physiological processes that lead to both obesity and death through independent mechanisms. 'So what,' one could argue, 'even if the observed association between BMI and mortality were mostly the result of having 'bad genes,' would that not still be an interesting finding? We would have proven that obese people are more likely to die.' This may be a helpful finding if we are only looking to identify those at high risk of mortality. But, if the goal is to guide public health policy to *reduce* mortality, that finding would be of little utility. To be sure that reducing BMI would reduce mortality, one needs to believe the assumption of conditional exchangeability (no unmeasured confounding). Otherwise, efforts to reduce BMI that did not affect the processes leading to both obesity and death would not achieve the goal of reduced mortality. As explained above, however, identifying all relevant confounders, and thus the causal effect of interest, is even harder for ill-defined interventions than for well-defined ones.

Consistency and positivity

Suppose that we are able to achieve conditional exchangeability. The exposed ($A=1$) and the unexposed ($A=0$) are comparable within the strata defined by the confounders L. A second condition for causal inference is that there are some exposed and some unexposed subjects in each stratum of L in the population. Equivalently, the probability of finding subjects with all levels of exposure must be *positive* (that is, greater than zero) in each stratum. We refer to this condition as the positivity condition.

Now, let us turn our attention to the observational study with the exposure BMI and the outcome mortality under the assumption that all confounders L are known and have been measured. Those confounders will include all genetic factors, or their proxies, that are joint determinants of body weight and mortality. But it is possible that some genetic traits exert such a tight control on body weight that, in the absence of malnutrition, all subjects possessing them will have a BMI >22. If these genetic traits also affect mortality, then they are confounders and should be adjusted for in the analysis. However, within some strata defined by the confounders, no subjects will have a BMI of 22; that is, positivity does not hold.

One potential solution to preserve positivity in this observational study would be to restrict the analysis to the strata of L in which the population contains some subjects with a BMI of 22 and others with a BMI of 30. With enough knowledge of the processes that determine BMI, one could restrict the analysis to people without the genetic trait that ensures a BMI $>$ 22. The price to pay for this strategy is a potential lack of generalizability of the estimated effect, as the subset of the population included in the analysis is no longer representative of the population of the country. This lack of generalizability is a commonly criticized feature of randomized experiments with strict eligibility criteria.

The violation of positivity points to another potential problem of ill-defined interventions: they may be unreasonable. The apparently straightforward comparison of subjects with a BMI of 22 and a BMI of 30 in observational studies masks the contrast of the implicit interventions 'make everybody in the population have a BMI of 22' and 'make everybody in the population have a BMI of 30.' Had these interventions been made explicit, investigators would have realized that the interventions were too extreme to be relevant from a public health standpoint. In fact, because drastic changes in BMI (say, from 30 to 22) in a short period of time may be unachievable and therefore unobserved in the data, any estimate of the effect of that intervention will rely heavily on modeling assumptions. A more reasonable, even if still ill-defined, intervention may be to reduce BMI by 10%.

In summary, violations of positivity are more likely to occur when estimating the effect of extreme interventions, and extreme interventions are more likely to go unrecognized when they are ill defined.

Conclusions

In an observational study, one can easily compare the mortality of those with BMI in the obesity range and of those with BMI in the normal range. At first glance, one might naively think that this contrast is sufficient to identify

'the' causal effect of BMI on mortality. However, because subjects achieve their BMI through some combination of mechanisms (diet, exercise, genes, illness and so on), this contrast implicitly considers a combined intervention on those mechanisms. Specifically, it considers the intervention 'assign everybody to normal weight by changing the determinants of BMI to reflect the distribution of those determinants in those who already have normal weight in the population.' This intervention is not straightforward. Furthermore, if the observational study adjusts for confounders, as it should, the implied intervention is then conditional on those confounders. If we adjust for diet, for example, the implied intervention is thus one in which the other nondiet determinants of BMI are changed to reflect the distribution of those determinants in those with normal BMI. Thus, the better we are able to control for known, measurable factors that determine both BMI and mortality, such as diet, exercise, cigarette smoking and illness, the more we are isolating an implied intervention that changes the remaining determinants of BMI, likely genes and physiology. Given a goal of informing policy, it would then seem that we have strayed from the most interesting questions.

Cigarette smoking offers an analogy: There are many pathways—such as DNA damage and oxidative stress—through which cigarette smoking affects mortality. However, one rarely (if ever) tries to estimate the excess deaths caused by intermediate variables, such as DNA damage or oxidative stress. They are not considered interesting estimates from a policy point of view because the relevant interventions are not clear. Now, replace cigarette smoking by physical activity and DNA damage by body weight. Perhaps because weight is so easy to measure, researchers seem to be interested in estimating the effect of weight changes without specifying the interventions that lead to those changes. Unfortunately, this approach does not necessarily lead to a meaningful causal inference.

Causal effects cannot be defined, much less computed, in the absence of a well-defined intervention.[8,10] As we have seen, one way to think of the problem is that without a well-defined intervention, we can have a gross violation of the consistency assumption. Different methods to change BMI may lead to different counterfactual outcomes, even if they each achieve the same BMI value. Thus the counterfactual outcome at a given BMI value cannot be defined. Without well-defined counterfactual outcomes, the assumption of conditional exchangeability or no unmeasured confounding becomes even less likely to be met than usual. Attempting to control all confounding requires delving so far into the biological processes that we may encounter violations of the positivity assumption.

The concerns about consistency raised in this paper are not unique to interventions on weight. All causal effects from observational data are vague, but there is a question of degree. For example, the effects of the intervention '1 h of daily strenuous exercise' may depend on how that hour would otherwise be spent. Reducing time spent laughing with friends, playing with your children or rehearsing with your band may have a different effect on mortality than reducing time eating, watching television or studying. However, the vagueness inherent in this intervention is relatively minor compared with the vagueness associated with BMI and can be reduced by a more careful specification of the intervention. The effects of exposures that are the result of complex biological processes—such as BMI, low-density lipoprotein-cholesterol, CD4 cell count or C-reactive protein—are vague because different methods of changing the exposure (diet, exercise, cigarette smoking and so on) may themselves have strong effects on mortality. The vagueness inherent in such exposures has led some authors to propose that only the causal effects of exposures that can be manipulated should be estimated (no causation without manipulation).[11] Estimating the 'effect' of exposures such as BMI may be an interesting hypothesis-generating exercise, but it is unclear whether the estimates can be directly translated into policy.

The ambiguities generated by violations of consistency cannot be dealt with by applying sophisticated statistical methods. All analytic methods for causal inference from observational data (stratification/regression, matching, inverse probability weighting, G-estimation, instrumental variable methods and so on) yield effect estimates that are only as well defined as the interventions that are being compared. Although the positivity condition can be waived if one is willing to make modeling assumptions to extrapolate to conditions that are not observed in the population, the consistency condition is so fundamental that it cannot be waived without simultaneously waiving the possibility of unambiguously describing the causal effect that is being estimated.

There are several other important issues related to the identification of well-defined interventions that are not addressed in this paper. One of them is that the most realistic interventions often take place over time. Identifying the causal effects of such interventions may require specific statistical methods that appropriately handle time-varying exposures and adjust for time-varying confounders.[12] Also, the choice of the outcome is crucial for a well-defined intervention. For example, consider the outcome 'excess deaths.' At least one study[4] found that the number of deaths attributable to obesity is decreasing over time in the United States. This finding, which might be explained by improved health care for obesity-related morbidities, begs the question of whether 'excess deaths' is the most relevant metric to measure the public health impact of obesity. A metric that weighs deaths by age (for example, life-years lost) or by health status (for example, quality-adjusted life-years lost) may be more informative as deaths are never prevented, just delayed.

To conclude, if the goal is to inform policy, it may be better to focus on modifiable lifestyle behaviors than on obesity itself, regardless of whether we use observational studies or randomized trials. We have discussed above that the lack of randomization in observational studies makes it necessary to assume that no unmeasured confounding exists, a strong

uncheckable assumption. However, if one is willing to believe that the assumption of no unmeasured confounding is approximately true, observational data can be used to mimic randomized experiments in which subjects or populations are 'assigned' to a well-defined intervention, followed for a certain time, and the outcome distribution compared among intervention groups. By considering observational studies in this way, we avoid tackling questions that cannot be logically asked in randomized experiments.

Acknowledgements

We thank Sander Greenland, Sonia Hernández-Díaz and Karen Steinberg for their detailed comments and expert advice. This work was supported by NIH Grant R01 HL080644.

Conflict of interest

Neither author declared any financial interests.

References

1 Allison DB, Fontaine KR, Manson JE, Stevens J, VanItallie TB. Annual deaths attributable to obesity in the United States. *J Am Med Assoc* 1999; **282**: 1530–1538.
2 Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. *J Am Med Assoc* 2004; **291**: 1238–1245.
3 Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Correction: actual causes of death in the United States, 2000. *J Am Med Assoc* 2005; **293**: 293–294.
4 Flegal KM, Graubard BI, Williamson DF, Gail MH. Excess deaths associated with underweight, overweight, and obesity. *J Am Med Assoc* 2005; **293**: 1861–1867.
5 Hernán MA. A definition of causal effect for epidemiological research. *J Epidemiol Commun Health* 2004; **58**: 265–271.
6 Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Commun Health* 2006; **60**: 578–586.
7 Robins JM, Greenland S. Comment on 'Causal inference without counterfactuals' by AP Dawid. *J Am Stat Assoc* 2000; **95**: 431–435.
8 Hernán MA. Invited commentary: hypothetical interventions to define causal effects: afterthought or prerequisite? *Am J Epidemiol* 2005; **162**: 618–620.
9 Greenland S, Rothman KJ. Measures of effect and measures of association. In: Rothman KJ, Greenland S (eds). *Modern Epidemiology*, 2nd edn. Lippincott-Raven: Philadelphia, 1998, pp 47–64.
10 Greenland S. Epidemiologic measures and policy formulation: lessons from potential outcomes (with discussion). *Emerging Themes Epidemiol* 2005; **2**: 5.
11 Holland PW. Statistics and causal inference. *J Am Stat Assoc* 1986; **81**: 945–961.
12 Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G (eds). *Advances in Longitudinal Data Analysis*. Chapman & Hall/CRC Press: New York, 2008 (in press).

Appendix

Formal definitions

We say that a counterfactual outcome Y_a is *consistent* with the actual or realized outcome Y if $Y_a = Y$ when the subject received exposure level $A = a$.

Exchangeability means that any counterfactual outcome under any treatment level a is independent of the treatment actually received A, which is written symbolically as $Y_a \perp\!\!\!\perp A$. Randomization of the treatment is expected to result in exchangeability. Stratified randomization, in which the probability of receiving treatment varies by levels of the variables in L, is expected to result in *conditional exchangeability* within levels L, which is written as $Y_a \perp\!\!\!\perp A|L$.

For discrete treatment A and covariates L, the *positivity* condition is written as $\Pr[A = a|L = l] > 0$ if $\Pr[L = l] \neq 0$. In general, positivity is written as $f_{A|L}(a|l) > 0$ if $f_L(l) \neq 0$, where $f_{X|Z}(x|z)$ is the conditional density function of the random variable X evaluated at the value x given the random variable Z evaluated at the value z.