Tyler J. VanderWeele* and Mirjam J. Knol

A Tutorial on Interaction

Abstract: In this tutorial, we provide a broad introduction to the topic of interaction between the effects of exposures. We discuss interaction on both additive and multiplicative scales using risks, and we discuss their relation to statistical models (e.g. linear, log-linear, and logistic models). We discuss and evaluate arguments that have been made for using additive or multiplicative scales to assess interaction. We further discuss approaches to presenting interaction analyses, different mechanistic forms of interaction, when interaction is robust to unmeasured confounding, interaction for continuous outcomes, qualitative or "crossover" interactions, methods for attributing effects to interactions, case-only estimators of interaction, and power and sample size calculations for additive and multiplicative interaction.

Keywords: effect modification, interaction, synergism, confounding, moderation

DOI 10.1515/em-2013-0005

It is not uncommon for the effect of one exposure on an outcome to depend in some way on the presence or absence of another exposure. When this is the case, we say that there is interaction between the two exposures. Recent years have seen increasing interest in interaction between genetic and environmental exposures, but interaction can also occur between two (or more) environmental exposures, or two different genetic exposures, or with various behavioral exposures. The processes giving rise to illness, health, and a variety of other outcomes are often inherently complex. Interaction between exposures is one manifestation of this complexity.

In this paper, we provide a tutorial on interaction. Many papers and book chapters discussing interaction are restricted to a fairly narrow set of issues. In this tutorial, we hope to provide a more comprehensive overview of issues related to interaction, primarily from the perspective of what has been written on the topic of interaction within the epidemiologic literature. However, we believe the tutorial will be of use for applied researchers throughout the biomedical and social sciences.

In this tutorial, we discuss the concept of interaction, some of the motivation for studying interaction, forms of statistical interaction and the issue of scale dependence, methods for estimating additive and multiplicative interaction, issues of confounding control and the causal interpretation of interaction measures, and how best to present interaction analyses. We also cover a number of more specialized topics including so-called "qualitative" or "crossover" interactions, interaction in the sufficient cause framework and in other mechanistic senses, the limits of statistical inference about biologic or physical interactions, methods for attributing effects to interactions, case-only designs for interaction, interaction for continuous outcomes, methods to identify subgroups to target using multiple covariates, the role of unmeasured confounding in interaction analyses, and power and sample size calculations for interac-

***Corresponding author: Tyler J. VanderWeele,** Departments of Epidemiology and Biostatistics, Harvard University, 677 Huntington Avenue, Boston, MA 02138, USA, E-mail: tvanderw@hsph.harvard.edu

Mirjam J. Knol, National Institute for Public Health and the Environment, RIVM, Bilthoven, The Netherlands, E-mail: mirjam.knol@rivm.nl

tion. The tutorial is long and is perhaps best read in two separate sittings. We have divided the tutorial into two parts: "Part I: Fundamental Concepts and Approaches for Interaction" and "Part II: Limitations, Extensions, Study Design, and Properties of Interaction Analysis" Part I is more introductory and accessible; Part II covers some more advanced topics and some of these are a bit more technical.

1 Part I: Fundamental concepts and approaches for interaction

1.1 Motivations for assessing interaction

There are a number of practical and theoretical considerations that motivate the study of interaction. One of the most prominent of these is that, in a number of settings, resources to implement interventions may be limited. It may not be possible to intervene on or treat an entire population. Resources may only be sufficient to treat a small fraction. If this is the case, then it may be important to identify the subgroups of individuals in which the intervention or treatment is likely to have the largest effect. As will be discussed below, methods for assessing additive interaction can help determine which subgroups would benefit most from treatment. Other more sophisticated methods can help identify groups of individuals, based on a large number of covariates, who would or would not benefit, or who would benefit to the greatest extent, from treatment. Even in settings in which resources are not limited and it is possible to intervene on everyone, it may be the case that a particular intervention is beneficial for some individuals and harmful for others. In such cases, it is very important to identify those groups for which treatment may be harmful and refrain from treating such persons. Techniques for assessing such so-called "qualitative" or "crossover" interactions are discussed in this tutorial are useful in this regard.

Another reason sometimes given for empirically assessing interaction is that it may provide insight into the mechanisms for the outcome. We will describe in this tutorial how it is possible to sometimes detect individuals for whom an outcome would occur if both exposures are present but would not occur if just one or the other were present. We will see that this more mechanistic notion of interaction is quite distinct from more statistically-based notions of interaction; we will see that in some cases we can gain insight into whether there might be a mechanism requiring two or more specific causes to operate and we will discuss the limits of such reasoning. Yet another reason sometimes given for studying interaction is that leveraging interactions that may be present may in fact help increase power in testing for the overall effect of an exposure on an outcome. In some settings, by jointly testing for a main effect and for an interaction simultaneously, it is possible to detect an overall effect when tests that ignore the interaction would not otherwise be able to detect the effect. It has been proposed that this may be especially important in the context of studying genetic variants when many variants are being tested and correction for such multiple testing reduces power, whereas allowing for the joint test may increase power to detect the effects.

As noted above, one of the motivations for studying interaction is to identify which subgroups would benefit most from intervention when resources are limited. However, in some settings, it may not be possible to intervene directly on the primary exposure of interest, and one might instead be interested in which other covariates could be intervened upon to eliminate much or most of the effect of the primary exposure of interest. In these cases, methods for attributing effects to interactions, discussed in the latter part of the tutorial, can be useful in assessing this and identifying the

most relevant covariates for intervention. Finally, sometimes interactions are modeled not with any specific scientific or policy goal in mind concerning interactions per se, but simply because the statistical model fits the data better when the model includes the additional flexibility allowed by an interaction term. These various motivations for studying interaction are distinct and, as we will see throughout, when studying interaction it is important to clearly understand what the goal of the analysis is.

1.2 Measures of interaction and scale of interaction

As a motivating example, consider data presented in Hilt et al. (1986) concerning the effect of smoking on lung cancer and how this varied by previous exposure to asbestos. The risk of lung cancer comparing smokers and non-smokers varied by asbestos exposure as presented in Table 1.

Table 1 Risk of lung cancer by smoking and asbestos status

	No asbestos	Asbestos
Non-smoker	0.0011	0.0067
Smoker	0.0095	0.0450

It seems as though lung cancer risk is much higher when both smoking and asbestos exposure are present together. This is an example of what we might call an interaction.

Let D denote a binary outcome. Let G and E denote two binary exposures of interest. These might be a genetic factor and an environmental factor, respectively, but our discussion will not be restricted to gene−environment interaction and G and E could represent any two factors; later in the tutorial we will also discuss interaction when the factors are not binary, but much of the discussion here generalizes in a straightforward manner. Let $p_{ge} = P(D = 1|G = g, E = e)$ be the probability of the outcome when G is value g and E is value e. A natural way to assess interaction is to measure the extent to which the effect of the two factors together exceeds the effect of each considered individually (cf. Rothman, 1986; Szklo and Nieto, 2007). This could be measured by:

$$(p_{11} - p_{00}) - [(p_{10} - p_{00}) + (p_{01} - p_{00})]. \quad [1]$$

Here $(p_{11} - p_{00})$ would be interpreted as the effect of both factors together compared to the reference category of both factors absent. The expressions $(p_{10} - p_{00})$ and $(p_{01} - p_{00})$ would be the effects of the first factor alone and the second factor alone, respectively. We would then consider the contrast between the effects of both factors together versus the sum of each considered separately. If this difference were non-zero we might say that there was interaction on the difference scale. For now, we will assume that the probabilities of the outcome under different exposure combinations correspond to the actual effects of the exposures on the outcome; we will consider issues of confounding and covariate adjustment in interaction analyses further below.

The measure in eq. [1] is sometimes referred to as a measure of interaction on the additive scale. The measure in eq. [1] can be rewritten as:

$$p_{11} - p_{10} - p_{01} + p_{00}. \quad [2]$$

If $p_{11} - p_{10} - p_{01} + p_{00} > 0$, the interaction is sometimes said to be positive or "super-additive." If $p_{11} - p_{10} - p_{01} + p_{00} < 0$, the interaction is said to be negative or "sub-additive".

For the data in Table 1, we have

$$\begin{aligned} & p_{11} - p_{10} - p_{01} + p_{00} \\ & = 0.0450 - 0.0095 - 0.00670 + .0011 \\ & = 0.0299. \end{aligned}$$

We would have evidence here of positive or "super-additive" interaction.

Sometimes, instead of using risk differences to measure effects, one might use risk ratios or odds ratios. For example, we could define the risk ratio effect measures as:

$$RR_{10} = p_{10}/p_{00},$$

$$RR_{01} = p_{01}/p_{00},$$

$$RR_{11} = p_{11}/p_{00}.$$

A measure of interaction on the multiplicative scale for risk ratios could then be taken as:

$$\frac{RR_{11}}{RR_{10}RR_{01}} = \frac{p_{11}p_{00}}{p_{10}p_{01}}. \quad [3]$$

This quantity measures the extent to which, on the risk ratio scale, the effect of both exposures together exceeds the product of the effects of the two exposures considered separately. If $RR_{11}/(RR_{10}RR_{01}) > 1$, the multiplicative interaction is said to be positive. If $RR_{11}/(RR_{10}RR_{01}) < 1$, the multiplicative interaction is said to be negative. Note that we compare the measure $RR_{11}/(RR_{10}RR_{01})$ to 1 rather than to 0 here since $RR_{11}/(RR_{10}RR_{01})$ is a ratio. If the ratio is 1, then the effect of both exposures together is equal to the product of the effect of the two exposures considered separately, that is, there is no interaction on the multiplicative scale for risk ratios. This measure of multiplicative interaction can also be rewritten as $\frac{RR_{11}}{RR_{10}RR_{01}} = \frac{p_{11}}{p_{01}} / \frac{p_{10}}{p_{00}}$, i.e. as the ratio of (i) the relative risk for G when $E = 1$ versus (ii) the relative risk for G when $E = 0$. Likewise, it can be written as $\frac{RR_{11}}{RR_{10}RR_{01}} = \frac{p_{11}}{p_{10}} / \frac{p_{01}}{p_{00}}$, i.e. as the ratio of (i) the relative risk for E when $G = 1$ versus (ii) the relative risk for E when $G = 0$.

Using the data in Table 1, we have that the measure of multiplicative interaction is given by:

$$\begin{aligned} & \frac{RR_{11}}{RR_{10}RR_{01}} \\ & = \frac{(0.0450/0.0011)}{\{(0.0095/0.0011) \times (0.0067/0.0011)\}} \\ & = \frac{40.9}{8.6 \times 6.1} = 0.78. \end{aligned}$$

We would have evidence here of negative multiplicative interaction.

This example also demonstrates that whether an interaction is positive or negative may depend on the scale. We may have a positive interaction on the additive scale but a negative interaction on a multiplicative scale. Said another way, the effect of both exposures together on the risk difference scale may exceed the sum of the effects on the risk difference scale of each considered separately, while it also being the case that the risk ratio for both exposures together is less than the product of the effects of the two exposures considered separately.

Likewise, interaction may be present on one scale but absent on another. Consider the data in Table 2.

Table 2 Risk of outcome by cross-classified exposure status

	$E = 0$	$E = 1$
$G = 0$	0.02	0.05
$G = 1$	0.07	0.10

Here there is no additive interaction since $p_{11} - p_{10} - p_{01} + p_{00} = 0.10 - 0.07 - 0.05 + 0.02 = 0$ but there is a negative multiplicative interaction since $RR_{11}/(RR_{10}RR_{01}) = (0.10/0.02)/\{(0.07/0.02)(0.05/0.02)\} = 5/(3.5 \times 2.5) = 0.57 < 1$. Likewise in other settings we might have additive interaction but no multiplicative interaction. Consider the data in Table 3.

Table 3 Risk of outcome by cross-classified exposure status

	$E=0$	$E=1$
$G=0$	0.02	0.05
$G=1$	0.04	0.10

Here the additive interaction is positive since $p_{11} - p_{10} - p_{01} + p_{00} = 0.10 - 0.04 - 0.05 + 0.02 = 0.03 > 0$, but there is no multiplicative interaction since $RR_{11}/(RR_{10}RR_{01}) = (0.10/0.02)/\{(0.04/0.02)(0.05/0.02)\} = 5/(2 \times 2.5) = 1$. In fact it can be shown (cf. Greenland et al., 2008) that if both of the two exposures have an effect on the outcome, then the absence of interaction on the additive scale implies the presence of multiplicative interaction for relative risks and likewise, the absence of multiplicative interaction for relative risks implies the presence of additive interaction. In other words, if both of the two exposures have an effect on the outcome, then there must be interaction on some scale. This raises the question of why interaction is of interest and which scale is to be preferred. It also makes clear that just to say that there is an interaction on some scale is relatively uninteresting; all it means is that both exposures have some effect on the outcome. Once again, when undertaking interaction analyses it is important to clarify what the goal or the motivation for the analysis is and choose a measure of interaction accordingly. In a subsequent section, we will turn to the arguments for and interpretation of additive versus multiplicative interaction. In general, however, either the presence or the absence of additive or multiplicative interaction may be of interest, and so it may be good practice to evaluate both additive and multiplicative interactions.

One reason why additive interaction is important to assess (rather than only relying on multiplicative interaction measures) is that it is the more relevant public health measure (Blot and Day, 1979; Saracci, 1980; Rothman et al., 1980; Greenland et al., 2008). Consider again the outcome probabilities in Table 3. Suppose that the outcome probabilities represent the probability of a disease being cured for a drug (E) stratified by genotype status (G). The effect of E on the risk difference scale among those with $G = 0$ is $0.05 - 0.02 = 0.03$; while the effect of E among those with $G = 1$ is $0.10 - 0.04 = 0.06$. If we had only 100 doses of the drug and we had to decide which group to treat, we could cure three additional persons if we used all of the drug supply among those with $G = 0$, but we could cure six additional persons if we used all of the drug supply among those with $G = 1$. All other things being equal, we would clearly want to give the drug supply to those with $G = 1$. The additive interaction measure, $p_{11} - p_{10} - p_{01} + p_{00} = 0.03 > 0$, allows us to see this. The multiplicative interaction measure, $RR_{11}/(RR_{10}RR_{01}) = 1$, does not.

In fact, the multiplicative scale can indicate the wrong subgroup to treat. Suppose in Table 3 we replace the final probability of cure 0.10 with 0.09. Then the effect on the difference scale of E among those with $G = 0$ is $0.05 - 0.02 = 0.03$; the effect of E among those with $G = 1$ is $0.09 - 0.04 = 0.05$. Thus, on the difference scale, the effect size is larger for the $G = 1$ subgroup, indicating this is the subgroup we would like to treat if resources are limited. However, on the risk ratio scale, the effect for those with $G = 0$ is $0.05/0.02 = 2.5$ and for those with $G = 1$ it is $0.09/0.04 = 2.25$; the risk ratio effect size is larger for the $G = 0$ subgroup; however, this is not the subgroup we would want to allocate limited resources to. If we had only 100 doses of the drug, we could cure three additional persons if we used all of the drug supply among those with $G = 0$, but we could cure five additional persons if we used all of the drug supply among those with $G = 1$. All other things being equal, we would clearly want to give the drug supply to those with $G = 1$. The issue with the multiplicative scale here is that the baseline risk is different in the two subgroups, and thus the risk ratio is operating on different baseline risks.

The possibility of positive additive interaction but negative or null multiplicative interaction is not simply a theoretical possibility. This was precisely the situation with the lung cancer data in Table 1 where

we had a positive additive interaction, but a negative multiplicative interaction. It was likewise the case in analyses of the joint effects of *Helicobacter pylori* and use of NSAIDs in causing peptic ulcer (Kuyvenhoven et al., 1999) with slightly positive additive interaction but negative multiplicative interaction. Similarly, in analyses of interaction between factor V Leiden mutation and oral contraceptive use in causing venous thrombosis, the multiplicative interaction was found to be close to null, but there was a positive additive interaction (Vandenbroucke et al., 1994). Using the multiplicative interaction results in any of these cases to determine which subgroups to prioritize intervention would have given the wrong conclusion. For example, from the data in Table 1 more lives would be saved by removing asbestos from homes of smokers first; the risk ratios give the opposite conclusion. Indeed dismissing the importance of one factor in assessing the effects of another because of the absence of multiplicative interaction can be quite dangerous: the null multiplicative interaction between factor V Leiden mutation and oral contraceptive use may lead to false reassurances that "it does not matter" whether one carries the mutation or not for the decision to start using oral contraceptives; whereas, in fact, because those with the factor V Leiden mutation have a roughly seven times higher baseline risk than those without the mutation (Vandenbroucke et al., 1994), the "constant risk ratio" for oral contraceptive use results in a much higher increase in absolute risk for those with the factor V Leiden mutation than those without.

More generally, $p_{11} - p_{10} - p_{01} + p_{00} > 0$ implies the public health consequence of an intervention on E would be larger in the $G = 1$ group, while $p_{11} - p_{10} - p_{01} + p_{00} < 0$ implies the public health consequence of an intervention on E would be larger in the $G = 0$ group. Thus, while it may be of interest to assess multiplicative interaction, additive interaction should also in general be examined, if for no other reason than to assess public health relevance.

In some case–control study designs, only the odds ratio can be evaluated and thus effect measures and interaction measures are evaluated on an odds ratio scale. The effects for each of the exposures considered separately and both considered together on the odds ratio scale are defined respectively by:

$$OR_{10} = \frac{p_{10}/(1-p_{10})}{p_{00}/(1-p_{00})},$$

$$OR_{01} = \frac{p_{01}/(1-p_{01})}{p_{00}/(1-p_{00})},$$

$$OR_{11} = \frac{p_{11}/(1-p_{11})}{p_{00}/(1-p_{00})}.$$

A measure of interaction on the multiplicative scale for odds ratio could then be taken as:

$$\frac{OR_{11}}{OR_{10}OR_{01}}. \qquad [4]$$

This quantity measures the extent to which, on the odds ratio scale, the effect of both exposures together exceeds the product of the effects of the two exposures considered separately. If $OR_{11}/(OR_{10}OR_{01}) > 1$, the multiplicative interaction is said to be positive. If $OR_{11}/(OR_{10}OR_{01}) < 1$, the interaction is said to be negative. For the data in Table 1, we have

$$\begin{aligned}\frac{OR_{11}}{OR_{10}OR_{01}} &= \frac{42.79}{8.71 \times 6.13} \\ &= 0.80.\end{aligned}$$

The measure of multiplicative interaction on the odds ratio scale is negative. The measure is very close to what was obtained for the multiplicative interaction on the risk ratio scale, i.e. 0.78.

In general, measures of multiplicative interaction on the odds ratio and risk ratio scales will be very close to one another whenever the outcome is rare. When the outcome is rare, both $(1-p_{ge})$ and $(1-p_{00})$ will be close to 1 and thus the odds ratios approximate risk ratios since

$$OR_{ge} = \frac{p_{ge}/(1-p_{ge})}{p_{00}/(1-p_{00})} \approx \frac{p_{ge}}{p_{00}} = RR_{ge}.$$

Odds ratios will also equal risk ratios (even when the outcome is common) in certain case–control designs in which the controls are selected from the entirety of the underlying population rather than just from the non-cases (cf., e.g. Knol et al., 2008, for further review and discussion of this point).

We may also be interested in assessing additive interaction from data when only relative risks are available or reported. Although we may not be able to estimate the additive interaction in eq. [2], i.e. $p_{11}-p_{10}-p_{01}+p_{00}$, directly, we can still proceed as follows. If we divide eq. [2] by p_{00} we obtain the following:

$$RERI_{RR} = RR_{11} - RR_{10} - RR_{01} + 1. \quad [5]$$

This quantity is sometimes referred to as the "relative excess risk due to interaction" or *RERI* (Rothman, 1986). It is also sometimes referred to as the "interaction contrast ratio" or ICR (Greenland et al., 2008). This gives us something similar to additive interaction but using risk ratios rather than risks. Subsequently, we will refer to this quantity in eq. [5] as $RERI_{RR}$. We have that $RERI_{RR} > 0$ if and only if for the additive interaction in eq. [2], $p_{11}-p_{10}-p_{01}+p_{00} > 0$; likewise $RERI_{RR} < 0$ if and only if $p_{11}-p_{10}-p_{01}+p_{00} < 0$; and $RERI_{RR} = 0$ if and only if $p_{11}-p_{10}-p_{01}+p_{00} = 0$. Thus, we can assess whether additive interaction is positive, negative, or zero using risk ratios and $RERI_{RR}$. It should be noted that although $RERI_{RR}$ gives the direction (positive, negative, or zero) of the additive interaction, we cannot in general use $RERI_{RR}$ to make statements about the relative magnitude of the underlying additive interaction for risks, $p_{11}-p_{10}-p_{01}+p_{00}$, unless we know p_{00}. We may have $RERI_{RR}$ larger in one of two subpopulations, but the additive interaction for risks, $p_{11}-p_{10}-p_{01}+p_{00}$, may be larger in the other; this is because the baseline risks, p_{00}, may differ and $RERI_{RR}$ depends on the baseline risk (Skrondal, 2003).[1] However, again, only the direction, rather than the magnitude, of $RERI_{RR}$ is needed to draw conclusions about the public health relevance of interaction. If we are trying to decide which subgroup of G to target for an intervention when resources are limited, $RERI_{RR} > 0$ implies the public health consequences of an intervention on E would be larger in the $G = 1$ group, while $RERI_{RR} < 0$ implies the public health consequences of an intervention on E would be larger in the $G = 0$ group.

A few other measures of additive interaction using data from risk ratios or odds ratios are sometimes employed. The so-called synergy index (Rothman, 1986) is defined as:

$$S = \frac{RR_{11}-1}{(RR_{10}-1)+(RR_{01}-1)}.$$

It measures the extent to which the risk ratio for both exposures together exceeds 1, and whether this is greater than the sum of the extent to which each of the risk ratios considered separately each exceed 1. Suppose the denominator of S is positive, then if $S > 1$ then we will have $RERI_{RR} > 0$ and thus

1 For example, suppose that the risks for G and E, stratified by gender, are: for males, $p_{00} = 0.02$, $p_{01} = 0.03$, $p_{10} = 0.03$, and $p_{11} = 0.06$ and for females $p_{00} = 0.01$, $p_{01} = 0.02$, $p_{10} = 0.02$, and $p_{11} = 0.05$. Then the additive interaction for risks for males is $p_{11}-p_{10}-p_{01}+p_{00} = 0.02$ and for females it is also $p_{11}-p_{10}-p_{01}+p_{00} = 0.02$. However if we examine $RERI_{RR}$ for males we get $RERI_{RR} = (p_{11}-p_{10}-p_{01}+p_{00})/p_{00} = 1$ but for females we obtain $RERI_{RR} = (p_{11}-p_{10}-p_{01}+p_{00})/p_{00} = 2$. We have a higher $RERI_{RR}$ for females than for males even though the underlying additive interaction for risks is the same. We obtain a higher $RERI_{RR}$ for females because the baseline risk for females $p_{00} = 0.01$ is lower than for males, $p_{00} = 0.02$. Again $RERI_{RR}$ can be used to assess the direction (positive, negative, or zero) of the additive interaction for risks but not the magnitude of the additive interaction for risks. If the magnitude (rather than just the sign) of $RERI_{RR}$ is going to be interpreted then it must be kept in mind that this magnitude is on the excess relative risk scale, and this does not necessarily correspond to the relative magnitude of additive interaction for risks. Once again, this is because the baseline risks may differ across groups.

$p_{11} - p_{10} - p_{01} + p_{00} > 0$; and if $S < 1$ then we will have $RERI_{RR} < 0$ and thus $p_{11} - p_{10} - p_{01} + p_{00} < 0$. Thus, the synergy index can likewise be used to assess additive interaction. The interpretation of the synergy index becomes difficult in settings in which one or both of the exposures is preventive rather than causative so that the denominator of S is negative (Knol et al., 2011).[2] This issue does not arise with $RERI_{RR}$ because the denominator of $RERI_{RR}$ is never negative. The issue can be resolved with the synergy index S by recoding the exposures so that neither is preventive in the absence of the other (Knol et al., 2011). Another measure of additive interaction that is sometimes used is called the attributable proportion and is defined as:

$$AP = \frac{RR_{11} - RR_{10} - RR_{01} + 1}{RR_{11}}$$

and essentially measures the proportion of the risk in the doubly exposed group that is due to the interaction itself. The attributable proportion is essentially a derivative measure of the relative excess risk due to interaction: $AP > 0$ if and only if $RERI_{RR} > 0$; and $AP < 0$ if and only if $RERI_{RR} < 0$. A variant on the attributable proportion may also be potentially of interest. The attributable proportion measured above, $AP = \frac{RERI_{RR}}{RR_{11}} = \frac{RR_{11} - RR_{10} - RR_{01} + 1}{RR_{11}} = \frac{p_{11} - p_{10} - p_{01} + p_{00}}{p_{11}}$, essentially measures the proportion of *risk* in the doubly exposed group that is due to interaction. Alternatively, we might consider the proportion of the joint *effects* of both exposures together that is due to interaction (Rothman, 1986; VanderWeele, 2013). This measure is given by $AP^* = \frac{RERI_{RR}}{RR_{11} - 1} = \frac{RR_{11} - RR_{10} - RR_{01} + 1}{RR_{11} - 1} = \frac{p_{11} - p_{10} - p_{01} + p_{00}}{p_{11} - p_{00}}$. Its properties will be considered later in the tutorial in the section on attributing effects to interactions.

All of these measures can be used in cohort studies, but these measures are also of interest and can be employed in case–control studies as well. Suppose that we only have estimates for odds ratios but that the outcome is rare (or that the controls are selected from the entirety of the underlying population rather than just from the non-cases cf. Knol et al., 2008) so that odds ratios approximate risk ratios. We could then replace each of the risk ratios in $RERI_{RR}$, the synergy index S, or the attributable proportion measures, with odds ratios to obtain approximations to each of these measures of additive interaction. For example, for the relative excess risk due to interaction, we can define $RERI_{OR} = OR_{11} - OR_{10} - OR_{01} + 1$, which is the odds ratio analog of $RERI_{RR}$. If the outcome is rare then we have that

$$\begin{aligned} RERI_{OR} &= OR_{11} - OR_{10} - OR_{01} + 1 \\ &\approx RR_{11} - RR_{10} - RR_{01} + 1 = RERI_{RR}. \end{aligned}$$

Thus, when odds ratio approximate risk ratios, we can assess additive interaction, at least approximately, even if only estimates of odds ratios are available from case–control study designs. Note that for this argument to apply using the assumption of a rare outcome (10% is often used as a threshold in practice), the outcome must be rare in each stratum defined by the two exposures. Sampling controls for the entire underlying population rather than only the non-cases removes the need for this rare outcome assumption.

As an example, Figueiredo et al. (2004) studied the effects of XRCC3-T241M polymorphisms and alcohol consumption on breast cancer risk using a case–control study design. The genetic risk factor was considered the M/M genotype versus a reference of the T/T or T/M genotype. They obtained the odds ratio in Table 4 from their case–control study.

2 When one or both of the exposures is preventive, rather than causative (i.e. $RR_{10} < 1$ and/or $RR_{01} < 1$), such that the denominator of S, $(RR_{10} - 1) + (RR_{01} - 1)$, is less than 0, then with an inequality like $S > 1$, multiplying both sides of this inequality by $(RR_{10} - 1) + (RR_{01} - 1)$, which is negative, will reverse the sign of the inequality, because of multiplication by a negative number, to give $RR_{11} - 1 < (RR_{10} - 1) + (RR_{01} - 1)$ or $RERI_{RR} < 0$; and thus when the denominator of S is negative, $S < 1$ becomes the condition for positive additive interaction, which can be confusing. In general, it is thus best not to report S unless the denominator, $(RR_{10} - 1) + (RR_{01} - 1)$, is positive.

Table 4 Odds ratios for breast cancer by strata of alcohol consumption and XRCC3-T241M

	No alcohol	Alcohol
T/T or T/M	1	1.12
M/M	1.21	2.09

Although we cannot assess additive interaction directly using risks, $p_{11} - p_{10} - p_{01} + p_{00}$, from the odds ratios in Table 4, we can still estimate

$$\begin{aligned} RERI_{OR} &= OR_{11} - OR_{10} - OR_{01} + 1 \\ &= 2.09 - 1.21 - 1.12 + 1 = 0.76 > 0 \end{aligned}$$

and so we would have evidence of positive additive interaction. Breast cancer is a relatively rare outcome, and so odds ratios will closely approximate risk ratios in this study. Likewise, we could calculate the synergy index $S = \frac{RR_{11}-1}{(RR_{10}-1)+(RR_{01}-1)} = 3.30 > 1$, again indicating positive additive interaction. And we can calculate the proportion of risk in the doubly exposed group attributable to interaction, $AP = \frac{RR_{11}-RR_{10}-RR_{01}+1}{RR_{11}} = 36.4\%$ or the proportion of the joint effects of both exposures attributable to interaction, $AP^* = \frac{RR_{11}-RR_{10}-RR_{01}+1}{RR_{11}-1} = 69.7\%$.

1.3 Statistical interactions and statistical inference

In practice, interactions are often evaluated by using statistical models by including a product term for the two exposures in the model. A statistical model on the linear scale accommodating interaction might take the form:

$$P(D = 1|G = g, E = e) = \alpha_0 + \alpha_1 g + \alpha_2 e + \alpha_3 eg. \quad [6]$$

It can be verified under this model that $\alpha_0 = p_{00}$, $\alpha_1 = p_{10} - p_{00}$, $\alpha_2 = p_{01} - p_{00}$, and $\alpha_3 = p_{11} - p_{10} - p_{01} + p_{00}$. The coefficient α_3 is thus equal to our measure of additive interaction based on risks; for this reason, α_3 is sometimes referred to as a statistical interaction on the additive scale.

Similarly, one might use a log-linear model for risk ratios, including a product term:

$$\log\{P(D = 1|G = g, E = e)\} = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 eg. \quad [7]$$

Here we have that $e^{\beta_0} = p_{00}$, $e^{\beta_1} = RR_{10}$, $e^{\beta_2} = RR_{01}$, and $e^{\beta_3} = RR_{11}/(RR_{10}RR_{01})$. The so-called "main effects", β_1 and β_2, when exponentiated, simply give the risk ratios for each of the two exposures when each is considered alone. The coefficient β_3, when exponentiated, gives our measure for multiplicative interaction for risk ratios, $RR_{11}/(RR_{10}RR_{01})$. The coefficient β_3 is thus often referred to as a statistical interaction for a log-linear model. Likewise, one might use a logistic model for odds ratios, including a product term:

$$\text{logit}\{P(D = 1|G = g, E = e)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg. \quad [8]$$

Here we have that $e^{\gamma_0} = p_{00}/(1 - p_{00})$, $e^{\gamma_1} = OR_{10}$, $e^{\gamma_2} = OR_{01}$, and $e^{\gamma_3} = OR_{11}/(OR_{10}OR_{01})$. The main effects, γ_1 and γ_2, when exponentiated, simply give the odds ratios for each of the two exposures. The coefficient γ_3, when exponentiated, gives our measure for multiplicative interaction for odds ratios, $OR_{11}/(OR_{10}OR_{01})$. Thus, γ_3 is referred to as a statistical interaction for a logistic model. The equality $e^{\gamma_0} = p_{00}/(1 - p_{00})$ will only hold with cohort data. However, all the other equalities, $e^{\gamma_1} = OR_{10}$, $e^{\gamma_2} = OR_{01}$, and $e^{\gamma_3} = OR_{11}/(OR_{10}OR_{01})$, will hold for both cohort data and case–control data. We can thus assess both of the main effects of the exposure and the multiplicative interaction between the exposures on an odds ratio scale using case–control data.

When the outcome and both exposures are binary, and no further covariates are included, it is straightforward to fit these models to the data using standard software. The estimate and confidence intervals obtained by maximum likelihood estimation and given by such software for α_3 will constitute an estimate and confidence interval for the additive interaction $p_{11} - p_{10} - p_{01} + p_{00}$. The estimate and confidence intervals obtained by maximum likelihood estimation and given by such software for β_3 and γ_3, when exponentiated, will constitute an estimate and confidence interval for the multiplicative interaction on the risk ratio and odds ratio scales, respectively. Statistical inference for interaction is thus straightforward in these cases.

Often we may want to control for other covariates in models [6]–[8]. For example, we may want to fit the following analogous models which include an additional vector of covariates, C:

$$P(D = 1|G = g, E = e, C = c) = \alpha_0 + \alpha_1 g + \alpha_2 e + \alpha_3 eg + \alpha_4' c,$$

$$\log\{P(D = 1|G = g, E = e, C = c)\} = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 eg + \beta_4' c,$$

$$\text{logit}\{P(D = 1|G = g, E = e, C = c)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg + \gamma_4' c.$$

Unfortunately, the linear and log-linear models, when fit to data, will often run into convergence problems in the maximum likelihood algorithms used to fit the models, especially when there are continuous covariates in C, because the models do not ensure that the predicted probabilities lie between 0 and 1. The logistic model with covariates does not suffer from this problem. For this reason, the most common approach to assessing interaction in practice has become fitting the logistic model with covariates and assessing the estimate and confidence interval for the product term coefficient, γ_3, in this model. This approach is also popular because it can be implemented in a straightforward way with case–control data as well. The coefficient, γ_3, is an important and useful measure of interaction and proceeding with this strategy is recommended.

However, as discussed throughout this tutorial, it is also recommended that investigators assess additive interaction as well. This can be more challenging when covariates are in the model. Additional strategies to fit linear and log-linear models with covariates using data from cohort studies have been described elsewhere (cf. Yelland et al., 2011; Knol et al., 2012, for overviews of several different methods). In the next section, however, we will describe what has now become a fairly standard approach (Hosmer and Lemeshow, 1992) to estimating additive interaction, with covariate control, which consists of using a logistic regression with additional covariates and transforming the parameter estimates to obtain estimates and confidence intervals for the relative excess risk due to interaction (*RERI*).

1.4 Inference for additive interaction

Suppose the following model is fit to the data:

$$\text{logit}\{P(D = 1|G = g, E = e, C = c)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg + \gamma_4' c. \quad [9]$$

We then have that

$$\begin{aligned} RERI_{OR} &= OR_{11} - OR_{10} - OR_{01} + 1 \\ &= e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1. \end{aligned}$$

Thus, we can estimate a measure of additive interaction, $RERI_{OR}$, using the parameters of a logistic regression. This approach has the advantage that the logistic regression in eq. [9] can more easily be fit to data when there are continuous covariates than the corresponding linear or log-linear models for binary outcomes given in the previous section. This approach with logistic regression also has the advantage that it can be employed even with case–control data. Even with cohort data, if the outcome is rare, this

approach to additive interaction using $RERI_{OR}$ can often be helpful because the logistic regression model often fits data quite well and has fewer convergence issues than a linear or log-linear model for risk, as discussed above. The logistic regression model has the interesting implication that if the model is correctly specified so that the log odds are linear in the covariates C, then the $RERI_{OR}$ measure will also be constant across strata of the covariates. This approach to $RERI_{OR}$, as other modeling approaches, presupposes that the statistical model is correctly specified. We discuss below other modeling approaches for additive interaction that make different modeling assumptions.

Standard errors for $RERI_{OR}$, as estimated above, can be obtained using the delta method (Hosmer and Lemeshow, 1992). Software options are now available to estimate these standard errors (e.g. Lundberg et al., 1996; Andersson et al., 2005).[3] In the Appendix, we provide some simple SAS code to estimate $RERI_{OR}$ and its standard error using the delta method. We likewise describe how this can be done in Stata (cf. Ai and Norton, 2003; Norton et al., 2004). Finally, as an online supplement to this tutorial, we have provided an Excel spreadsheet that can be used in conjunction with standard output from logistic regression (output on parameter estimates and either the covariance or correlation estimates) using any software package. The current Excel spreadsheet offers somewhat more flexibility than previous versions of the spreadsheet in allowing for confidence intervals of any percentile.

The approach described above works well if the outcome is rare so that $RERI_{OR}$ approximates $RERI_{RR}$. If the outcome is common, $RERI_{OR}$ may not be an adequate measure of additive interaction. In such cases, for cohort data, one could estimate $RERI_{RR}$ by replacing the logistic model in eq. [9] with a log-linear model, though such log-linear models with continuous covariates C may not always converge; likewise an approach for risk ratios using modified Poisson, rather than logistic regression, has also been proposed that can be used with a common outcome (Zou, 2008). Alternatively, with cohort data with a common outcome, one may use a weighting approach to estimating additive interaction (VanderWeele et al., 2010). This approach models the relationship between the exposures and the covariates, rather than between the outcome and the covariates.

Our discussion thus far has focused on binary exposures. A similar approach can be used with categorical, ordinal, or continuous exposures. The logistic regression model above in eq. [9] could be fit to the data if the two exposures G and E were ordinal or continuous. However, when additive interaction is carried out for ordinal or continuous exposures using this approach based on logistic regression, two things must be kept in mind, one analytical and one interpretative. First analytically, for ordinal and continuous exposures, it is important to consider the magnitude of the change in the exposures for which one is examining interaction. If one is considering a change for the value of G from g_0 to g_1 and a value of E from e_0 to e_1 then instead of using $e^{\gamma_1+\gamma_2+\gamma_3} - e^{\gamma_1} - e^{\gamma_2} + 1$ as an estimate of $RERI_{OR}$ one uses

$$RERI_{OR} = e^{(g_1-g_0)\gamma_1+(e_1-e_0)\gamma_2+(g_1e_1-g_0e_0)\gamma_3} - e^{(g_1-g_0)\gamma_1+(g_1-g_0)e_0\gamma_3} - e^{(e_1-e_0)\gamma_2+(e_1-e_0)g_0\gamma_3} + 1.$$

This needs to be taken into account when using the software and Excel spreadsheets, so that estimates and covariance matrices are multiplied by the appropriate factors. This is described in more detail in the

3 To estimate standard errors for $RERI_{OR}$ using logistic regression, in addition to the delta method described by Hosmer and Lemeshow (1992) and implemented with SAS and Stata code in the Appendix, one may also use bootstrapping which can have more accurate standard errors when the sample size is small (Assmann et al., 1996); other re-sampling based approaches are available when some of the outcome counts for particular exposure combinations are low (Nie et al., 2010). Bayesian approaches to $RERI_{OR}$ are also now available (Chu et al., 2011). When sample sizes are relatively large, the approaches to estimating $RERI_{OR}$ will give fairly comparable confidence intervals; when sample sizes are small the resampling-based approach may be more accurate. However, in general, fairly large sample sizes are required to detect interaction; thus, for the most part, in those very settings in which it is possible and reasonable to test for interaction, the various approaches to estimate $RERI_{OR}$ are likely to give comparable estimates and standard errors. We discuss issues of power and sample size further below. Easy to implement software (Richardson and Kaufman, 2009; Kuss et al., 2010) is also available for estimating $RERI_{OR}$ using so-called linear odds models (cf. Skrondal et al., 2003). This approach, however, can have difficulty handling continuous covariates C. Such covariates can be handled in linear odds models using a weighting approach for covariate control (VanderWeele and Vansteelandt, 2011), and this approach can be employed with case–control data as well.

Appendix. Similar expressions could be given using categorical exposures: under any specific statistical model and for any two levels of each of the two exposures, one simply calculates the three relative risks comparing the various exposure combinations to the reference group and one subtracts from the risk ratio of the doubly exposed group, the two risk ratios for each of the singly exposed groups and adds 1. The second, more interpretative point, when ordinal, continuous, or categorical exposures are being employed, is that it is important to keep in mind that the $RERI_{OR}$ measure (or the analogous $RERI_{RR}$ measure) does vary according to the levels being compared and can vary in sign as well. The additive interaction measure for a change in E from 10 to 20 and in G from 0 to 1 may be different than the additive interaction measure for a change in E from 20 to 30 and in G from 0 to 1, but again as noted above, the $RERI_{OR}$ measure should be interpreted as giving the direction of additive interaction (positive, negative, or zero) and its relative magnitude does not necessarily correspond to the relative magnitude of the additive interaction for absolute risks. See also Knol et al. (2007) for further discussion. SAS and Stata code are given in the Appendix.

1.5 Additive versus multiplicative interaction

The fact that interaction can be assessed on different scales and that interaction is scale dependent raises the question on which scale interaction should be assessed: additive or multiplicative or some other. The view of this tutorial is that it is almost always best to present both additive and multiplicative measures of interaction (Botto and Khoury, 2001; Vandenbroucke et al., 2007; Knol and VanderWeele, 2012). In practice, measures of multiplicative interaction, using logistic regression, are most frequently reported. This is very likely simply done because of convenience, rather than because careful thought has been given to which measure is to be preferred. Standard software using logistic regression will automatically give an estimate and confidence interval for multiplicative interaction. As noted in the previous section, additional work is required in most current software packages to obtain measures of additive interaction, and for this reason it is not often done. In a recent review of a random sample of 25 cohort and 50 case–control studies from the five most highly ranked epidemiological journals, Knol et al. (2009) noted that although 61% of the studies included at least as secondary analyses an assessment of effect modification or interaction, only one reported a measure of additive interaction. In our view, it is in general a mistake to not report additive interaction. As noted above and as discussed further below, additive interaction is always relevant for assessing the public health significance of an interaction. Although we believe both additive and multiplicative interactions should in general be reported, we nonetheless review some of the reasons that have been put forward for using one scale versus the other.

The difference scale is useful for assessing the public health importance of interventions and the public health significance of interaction (Blot and Day, 1979; Saracci, 1980; Rothman et al., 1980; Greenland et al., 2008). As noted above, if the effect of an intervention is larger on the difference scale in one subgroup versus another, then this indicates that there would be larger numbers for whom the disease was prevented/cured in giving a hundred individuals in the first subgroup treatment versus giving a hundred individuals in the second subgroup treatment. Such information is useful for targeting subpopulations for which the intervention is most effective. This will be relevant whenever resources are constrained and thus relevant also for cost-effectiveness (Greenland, 2009). As discussed above, the additive, not the multiplicative, scale gives this information. A second reason sometimes given for using additive interaction is that it more closely corresponds to tests for mechanistic interaction, rather than merely statistical interaction (Greenland et al., 2008; VanderWeele and Robins, 2007, 2008; VanderWeele, 2010a, 2010b). As discussed further below, tests for additive interaction can sometimes be used to detect synergism in Rothman's (1976) sufficient cause framework. Conceived of another way, assessing additive interaction can sometimes be used to assess whether there are persons for whom the outcome would occur if both exposures were present but not if only one or the other of the exposures were present. As discussed below, this ends up being a different, and in many cases stronger, notion of interaction than merely a statistical

interaction. We return to this point in later sections. Finally, as also discussed further below, tests for additive interaction are sometimes more powerful than tests for multiplicative interaction and thus for the purposes of discovery and detection, the additive scale may be preferred as well.

Several reasons are also often put forward for using the multiplicative scale. First, as noted above, it is easier to fit multiplicative models (such as logistic regression), and the multiplicative scale is the most natural scale on which to assess interaction for such models; moreover, when using such models, measures of multiplicative interaction are readily obtained from standard software. Second, it is sometimes claimed that there is in general less heterogeneity on the multiplicative scale. Studies of meta-analyses have suggested that in terms of statistical significance, the risk ratio and odds ratio are less heterogeneous than the risk difference (Engels et al., 2000; Sterne and Egger, 2001; Deeks and Altman, 2003).[4] However, it is not entirely clear the extent to which this is simply due to difference in power across the different scales or whether there is genuinely less heterogeneity. Nevertheless, if it is indeed the case that the multiplicative scales (odds ratio or risk ratio) are "less heterogeneous", and this indicates something about the underlying biology as to how effects typically operate (see comments on the "Limits of Biologic Inference" below), then detecting an interaction on a multiplicative scale may be of greater import than detecting interaction on the additive scale. A third reason sometimes given for using the multiplicative scale for overall effects (but also potentially applicable to interaction), stated in some epidemiology textbooks, is that the relative effect measures are better suited to "assessing causality". According to Poole (2010), this notion can be traced back to a paper by Cornfield et al. (1959) showing that smoking was strongly related to lung cancer but not to other diseases on a relative risk scale, while smoking seemed similarly related to lung cancer and also to other diseases on an absolute risk scale. Because specificity of effect was seen as a criterion of causality (Hill, 1965), the relative risk scale was seen as superior over the absolute risk scale in assessing causality. As noted by Poole (2010), whether the relative or absolute measure is more useful for "assessing causality" will, however, vary by setting. In some cases, such as that considered by Cornfield et al. (1959), the multiplicative scale may indeed prove to be more useful, and it might be thought that this general argument then is also relevant to interaction.

Arguments can be given in favor of each of the two scales. However, nothing prohibits investigators from reporting measures of interaction on both additive and multiplicative scales and, in most settings, we think this approach is the best because both can be informative (Botto and Khoury, 2001; Vandenbroucke et al., 2007; Knol and VanderWeele, 2012). The presence or absence of interaction on either scale may be of interest. However, as noted above, provided both exposures have an effect on the outcome, there will always be interaction on at least one scale.[5] The only way there can be no interaction on any scale is for one of the two exposures to have *no* effect on the outcome at all. Thus, the fact that interaction is present on some scale really is not of much interest; provided both exposures have an effect on the outcome, such interaction on some scale will *always* be present. This brings us back to the point that was made at the beginning of the tutorial, that, when studying interaction, it is important to clearly understand what the goal of the analysis is: What is it that we are trying to learn? What scientific or policy question are we trying to answer and how does an interaction analysis help us? We have seen above already that interaction on the additive scale gives insight into which subgroups are best to treat. We will see below that interaction on the additive scale can also sometimes give insight into more mechanistic forms of interaction. As also discussed below the absence of interaction on either the additive or the multiplicative scale may also give

4 Engels et al. (2000) found that for 107 of 125 meta-analyses (86%) the *p*-value for heterogeneity for risk differences was less than that for the odds ratios. With a *p*-value cutoff of 0.10, they found that 59 (47%) meta-analyses were heterogeneous for the risk difference and 44 (35%) were heterogeneous for the odds ratio. Deeks and Altman (2003) likewise reported that the risk difference was more heterogeneous than the odds ratio or risk ratio using 1,889 meta-analyses. Sterne and Egger (2001) reviewed 78 meta-analyses and found that the *p*-value for heterogeneity was less than 0.05 in 29%, 27%, and 35% of these meta-analyses, for the odds ratio, risk ratio, and risk difference, respectively.

5 Though there may not always be sufficient statistical power to detect it, a point we return to below.

some clues (though rarely definitive evidence) as to the underlying biology; likewise we will see that the presence of positive multiplicative interaction may give some clues as to mechanisms. But it is always important to clarify what the goal of the analysis is and what we are trying to learn. Again, the fact that there is interaction on some scale is otherwise nothing more than acknowledging that both exposures have some effect.

1.6 Confounding and the interpretation of interaction

Thus far, we have considered measures of interaction using risk differences, risk ratios, and odds ratios. In general, however, we want to know whether our effect estimates correspond to causal effects rather than mere associations. In observational studies, we thus attempt to control for confounding. Analytically, this is often done through regression adjustment for other covariates. In interaction analyses, we have two exposures and thus potentially two sets of confounding factors to consider. The causal interpretation of interaction measures depends on whether control has been made for one or both sets of confounding factors, or neither.

Suppose we have made control for one set of confounding factors, those for the relationship between our primary exposure of interest and the outcome, but that we have possibly not controlled for confounding of the relationship between the secondary factor defining subgroups and the outcome. We would in this case still be able to obtain valid estimates of the effect of the primary exposure within strata defined by our secondary factor. For example, suppose we found substantial interaction between a drug and hair color when examining some health outcome. If we had controlled for the confounding factors for the drug–outcome relationship, or if the drug were randomized, we could interpret our interaction measure as a measure of heterogeneity concerning how the actual causal effect of the drug varied across subgroups defined by hair color. If we found that the effect of our primary exposure varied by strata defined by the secondary factor in this way, then we might call this "effect heterogeneity" or "effect modification." This might be useful, for example, in decisions about which subpopulations to target in order to maximize the effect of interventions. Provided we have controlled for confounding of the relationship between the primary exposure and the outcome, these estimates of effect modification or effect heterogeneity could be useful even if we have not controlled for confounding of the relationship between the secondary factor and the outcome. What we would not know, however, is whether the effect heterogeneity is *due to* the secondary factor itself, or something else associated with it. If we have not controlled for confounding for the secondary factor, the secondary factor itself may simply be serving as a proxy for something that is causally relevant for the outcome (VanderWeele and Robins, 2007b). For example, if we found that the effect of the drug varied by strata defined by hair color, this may simply be due to the fact that hair color is associated with genotype and it is this that is causally relevant for modifying the effect of the drug on the outcome. If we were simply to dye someone's hair, this would not change the effect of the drug.

If we are interested principally in assessing the effect of the primary exposure within subgroups defined by a secondary factor then simply controlling for confounding for the relationship between the primary exposure and the outcome is sufficient. However, if we want to intervene on the secondary factor in order to change the effect of the primary exposure then we need to control for confounding of the relationships of both factors with the outcome. When we control for confounding for both factors we might refer to this as "causal interaction" in distinction from mere "effect heterogeneity" mentioned above (VanderWeele, 2009a).

As another example, VanderWeele and Knol (2011) considered a randomized trial for a housing intervention program for homeless adults to reduce the number of hospitalizations. Suppose that the effect of the housing program was examined within strata defined by whether the participants had at least part-time employment. Here, the housing program is randomized, but employment status is not. If it were found that the housing intervention had a larger effect for those with part-time employment than for those without, this could be used as a valid estimate for the effect of the intervention within these different subgroups and could be useful in subsequently targeting the intervention toward the subgroups for which it would be most

effective. By randomization, we have controlled for confounding for the housing intervention, but we have not necessarily controlled for confounding for employment status. Thus, while we could get valid estimates of effects of the housing intervention within strata defined by employment status, we could not draw conclusions on what would happen if we intervened on employment status as well to try to improve the effect of the intervention. Again, employment status has not been randomized. Employment status may, for instance, be serving as a proxy for mental health, and it may be that mental health is in fact what is relevant in altering the effects of the intervention. It is possible that if we intervened on employment status, without changing mental health, then this would not alter at all the effect of the housing intervention. We would only be able to assess what the effect of interventions on employment status in altering the effect of the housing intervention would be if we had controlled for confounding of the relationship between the factor defining subgroups, namely employment status, and the outcome.

In summary, if we are interested in identifying which subpopulations it is best to target with a particular intervention, then assessing effect heterogeneity is fine and only the confounding factors of the relation between exposure and outcome need be considered (though even here it is sometimes argued control for other factors can help with external validity and extrapolation to other settings). If we are interested in potentially intervening on the secondary factor to change the effects of the primary intervention (or if we are interested in assessing mechanistic interaction, described below), then we want measures of causal interaction and we would need to control for confounding for the relationships between both factors and the outcome.

In practice, typically a regression model is simply fit to the data, regressing the outcome on the two exposures, a product term, and possibly other covariates. However, whether the regression coefficient for the product term can be interpreted as a measure of effect heterogeneity or causal interaction or both or neither depends on what confounding factors have been controlled for. For effect heterogeneity, we only have one set of confounding factors to consider, just those for the relationship between the primary exposure and the outcome. For causal interaction, we have two sets of confounding factors to consider, those for the primary exposure and the outcome and those for the secondary factor and the outcome. Epidemiologists are careful to control for confounding and think carefully about confounding in observational studies for overall causal effects. However, too often issues of confounding have been neglected in interaction analyses. Careful thought needs to be given to interaction analyses in interpreting associations as causal and in distinguishing between whether attempt is being made to control for one or both sets of confounding factors; and which of "effect heterogeneity" (also sometimes called "effect modification") or "causal interaction" is of interest will depend upon the context.[6]

The terms "interaction" and "effect modification" in practice are often used interchangeably. In some sense, what we have called "effect modification" is still a type of interaction analysis; and what we have called "causal interaction" could almost be viewed as "effect modification" by intervening on a secondary variable (VanderWeele, 2009a, 2010c). There is some ambiguity in terminology and it would be difficult to insist on a particular set of rules for terminology. However, even if the terms themselves are used interchangeably, it is important to keep in mind that there are still two distinct concepts present. The distinction again has to do with whether one or two potential interventions are in view. Failure to take the distinction into account could lead to incorrect policy recommendations. In writing papers, researchers can make clear which of the two concepts is in view (without having to adopt a strict terminological stance) by clarifying, in a Methods section, whether confounding control is intended for one or both exposures, and by

6 Additional subtlies also arise in distinguishing between interaction and effect heterogeneity/modification. For example, VanderWeele (2009a) showed that there can be cases in which effect modification is present but not interaction; or when interaction is present but not effect modification. Likewise there are also cases in which effect modification measures are identified from the data, but interaction measures are not; there are more subtle cases in which interaction measures are identified from the data but effect modification measures are not. Finally, VanderWeele (2009a) also discussed how the analytic procedures required to fit marginal structural models (Robins et al., 2000) for effect modification/heterogeneity differ from those required to fit marginal structural models for interaction.

commenting, in a Discussion section, whether interventions on one or both exposures are being considered when interpreting the implications of the results.

1.7 Presenting interaction analyses

Careful thought should be given to the presentation of interaction analyses. Very often when interaction or effect modification is of interest, effect measures are presented for each stratum separately using separate reference groups. Suppose, for example, we had data as in Table 1 and that effect measures were computed on the risk ratio scale. We let $E = 1$ denote asbestos exposure and $E = 0$ the absence of asbestos exposure and we let $G = 1$ denote smoking and $G = 0$ non-smoking. It is not uncommon for papers to present, e.g. the (adjusted) risk ratio effect measures for say the exposure E separately across strata of the other factor G. For example, the effect measures might be presented as in Table 5.

Table 5 Risk ratios with separate reference groups (uninformative presentation)

	No asbestos ($E = 0$)	**Asbestos ($E = 1$)**
Non-smoker ($G = 0$)	1 (reference)	$RR = 6.09$
Smoker ($G = 1$)	1 (reference)	$RR = 4.74$

While this information can be useful to see that the risk ratio in the non-smoking ($G = 0$) stratum is larger than the risk ratio in the smoking ($G = 1$) stratum, and for calculating multiplicative interaction: $4.74/6.09 = 0.78$ as above, there are several other comparisons for which Table 5 is uninformative. For example, by presenting the analyses with separate reference groups (for each of the $G = 0$ and $G = 1$ strata), we will not know from such a presentation whether the $(G = 0, E = 1)$ subgroup or the $(G = 1, E = 0)$ subgroup is at higher risk for the outcome. In fact, simply from the information in Table 5, we would not know whether the $(G = 1, E = 1)$ subgroup or the $(G = 0, E = 1)$ subgroup is at higher risk for the outcome, or whether the $(G = 1, E = 0)$ subgroup or the $(G = 0, E = 0)$ subgroup is at higher risk for the outcome. Nor do we know from Table 5 what the sign is for measures of additive interaction. Because of these reasons current guidelines (Vandenbroucke et al., 2007; Knol and VanderWeele, 2012) recommend that interaction and effect modification analyses be presented with a single common reference group, say the $(G = 0, E = 0)$ subgroup, or that the original data be presented (Botto and Khoury, 2001). If risk ratios with a common reference group were used for the data in Table 1, the effects could then be presented in Table 6.

Table 6 Risk ratios with a common reference group (informative presentation)

	No asbestos ($E = 0$)	**Asbestos ($E = 1$)**
Non-smoker ($G = 0$)	1 (reference)	6.09
Smoker ($G = 1$)	8.64	40.91

From the information presented in Table 6, which uses a common reference group, we would know that the ordering of risk across $G \times E$ subgroups was $(G = 0, E = 0)$, then $(G = 0, E = 1)$, then $(G = 1, E = 0)$, and then $(G = 1, E = 1)$. We could still calculate the individual risk ratios for E in the different strata of G as: $6.09/1 = 6.09$ for $G = 0$ and $40.91/8.64 = 4.74$ for $G = 1$ (and we could also add these to the table if

desired). We could thus also estimate measures of multiplicative interaction. We could moreover estimate the risk ratios for G across strata of E: e.g. $8.64/1 = 8.64$ for $E = 0$ and $40.91/6.09 = 6.72$ for $E = 1$ (and we could present these in the table if desired). And we could moreover estimate measures of additive interaction from the information in Table 6: $RERI_{RR} = 40.91 - 8.64 - 6.09 + 1 = 27.18 > 0$. The presentation of interaction analyses in Table 6 thus gives the reader far more information (using a single common reference category) than the presentation in Table 5 (using multiple reference categories). Presenting interaction analyses using a common reference category such as the presentation in Table 6 is thus to be preferred. If the study is a cohort study then it may be even further preferable to present the actual risks, as in Table 1, in the cells of the table, rather than the risk ratios (Botto and Khoury, 2001; Knol and VanderWeele, 2012).

Knol and VanderWeele (2012) further suggested that when interaction and effect modification analyses are presented the following items all be given in a table: (1) risk differences or relative risks (or odds ratios if risk differences or relative risks cannot be calculated) for each (G, E) stratum with a single reference category (possibly taken as the stratum with the lowest risk of the outcome); (2) risk differences, relative risks, or odds ratios for G within strata of E, and for E within strata of G; (3) interaction measures on additive and multiplicative scales, along with confidence intervals and p-values for these; (4) the exposure-outcome confounders for which adjustment has been made either for one of the exposures (for effect modification/heterogeneity analyses) or for both of the exposures (for interaction analyses) with clear indication of whether attempt is being made to control for one or two sets of confounding factors. Knol and VanderWeele (2012) also considered different layout options for this information and how to further extend such presentations when one or both exposures has more than two levels. If multiple different interaction analyses are conducted in the same paper and presented in the same table, it may be desirable to put all of these items on a single line of a table so that multiple interactions analyses can be presented in the same table.

Careful thought should be given to presenting interaction analyses, so that the reader has the maximum amount of information available. In almost all cases, interaction analyses with a single reference group should be presented. Failure to do so will obscure information from the reader.

1.8 Qualitative interaction

In some cases, we might think that an exposure has a positive effect for one subgroup and a negative effect for a different subgroup. Such instances are sometimes referred as "qualitative interactions" or "crossover interactions" (Peto, 1982; Gail and Simon, 1985).[7] Unlike statistical interactions in which the effects within two subgroups are both in the same direction, but simply differ in magnitude, qualitative interactions do not depend on the scale that is being used (de González and Cox, 2007). If there is a qualitative interaction on the difference scale, there will also be a qualitative interaction on the ratio scale.

As an example of such qualitative interaction, Gail and Simon (1985) considered data from a trial of two therapies for breast cancer, one of which does and the other of which does not involve tamoxifen. For young patients under age 50 with low progesterone receptor levels, the treatment without tamoxifen led to higher proportions who were disease-free after 3 years. However, for all other groups (who were either older, or had higher progesterone receptor levels, or both) the treatment with tamoxifen led to higher proportions who were disease-free after 3 years. Here, we would likely want to give young patients with low progesterone receptor levels the treatment without tamoxifen and others the treatment with tamoxifen.

7 The term "quantitative interaction" is sometimes used exclusively for interactions which are not qualitative interactions (Peto, 1982). However, others use the term "quantitative interaction" to describe a statistical interaction on any scale, and prefer using "non-crossover interaction" for the presence of interaction which is not a "qualitative interaction" (Gail and Simon, 1985).

In an example like this, we see then that qualitative interaction is very important in decision-making. We discussed above that in settings in which the intervention is beneficial for everyone but the magnitude of the benefit varies across subgroups, additive interaction can be useful in assessing whether it would be better to target the intervention to some subgroups rather than others if resources are limited. However, in such settings, if resources are not limited and the intervention is beneficial for everyone we may well want to treat all subgroups. Qualitative interaction, in contrast, has implications for treatment or interventions decisions even if resources are unlimited. In the presence of qualitative interaction, we do not want to treat all subgroups, because the treatment is in fact harmful in some subgroups. If qualitative interaction is present, it is thus important to be able to detect it.[8]

Several statistical approaches have been developed for testing for such qualitative interaction (e.g. Gail and Simon, 1985; Piantadosi and Gail, 1993; Pan and Wolfe, 1997; Silvapulle, 2001; Li and Chan, 2006). The details of these various approaches and their power properties do vary, but they all essentially coincide when one is simply testing for qualitative interaction between two subgroups. The approaches differ when examining qualitative interaction across three or more subgroups.[9] When testing for qualitative interaction across two subgroups one particularly simple approach (Pan and Wolfe, 1997) to test for a qualitative interaction at the 5% significance level is to construct 90% confidence intervals for the exposure effect in each of the two subgroups. If, on a difference scale say, one of the 90% confidence intervals lies entirely above 0 and the other lies entirely below 0, then one would reject the null hypothesis of no qualitative interaction. Note that only 90% confidence intervals (not 95%) need to be constructed here. These confidence intervals will be narrower than the usual 95% confidence intervals. One could alternatively carry out the analysis on a ratio scale and construct 90% confidence intervals for the effects in each of the subgroups and examine whether one of these 90% confidence intervals was completely above 1 and whether the other was completely below 1.

A special case or limit case of qualitative interaction is what is sometimes called a pure interaction in which the exposure has no effect whatsoever in one subgroup but does have an effect in a different subgroup. Like qualitative interactions, pure interactions do not depend on the scale being used. An example of such a "pure" interaction might include certain genetic variants on chromosome 15q25.1 which seem to only affect lung cancer for individuals who smoke and otherwise appear to have no effect for those who do not smoke (Li et al., 2010). We will consider this example further below.

8 Often, in a randomized trial, if a particular treatment or drug is known to be detrimental in some subgroups, such subgroups are typically then excluded from the trial when choosing participants. If this is so, qualitative interaction would then not be apparent because the groups for which the treatment has harmful effects are excluded in advance.

9 The various approaches do differ when testing for qualitative interaction using more than two subgroups. Pan and Wolfe (1997) described a fairly straightforward way to carry out such testing. Their approach allows for multiple subgroups and allows also testing for qualitative interaction of at least a certain magnitude (rather than simply whether one of the effects is larger, and the other smaller, than 0); it essentially just requires constructing confidence intervals of various sizes depending on the number of subgroups. Their approach is equivalent to that described by Piantadosi and Gail (1993), sometimes referred to as the "range test," but the implementation described by Pan and Wolfe (1997) is easier to carry out. An alternative approach was proposed by Gail and Simon (1985) which involves not simply constructing confidence intervals for the effects in each subgroup but rather constructing a confidence interval for the sum of the positive versus negative standardized effects across subgroups. The approach of Gail and Simon (1985) tends to perform better when there are several subgroups with effects which are positive and several also with effects which are negative. The approaches of Piantadosi and Gail (1993) and Pan and Wolfe (1997) tend to perform better if the effects in most of the subgroups are in one direction and there are only one or very few subgroups for which the effect is in the opposite direction. The motivation for these various approaches involving several subgroups is often having a continuous covariate or multiple covariates of interest which might define subgroups for which a qualitative interaction is thought to be present. However, with continuous covariates or multiple covariates, an approach described later in this tutorial for detecting effect heterogeneity based on a vector of covariate values, and determining for which individuals the treatment effects are positive versus negative may ultimately prove to be more useful.

1.9 Synergism and mechanistic interactions

Thus far, we have been considering different notions of statistical interaction and their interpretation. We noted above that such notions of interaction were scale dependent. In this section, we will consider drawing conclusions about more mechanistic forms of interaction. We might say that a "sufficient cause interaction" is present, if there are individuals for whom the outcome would occur if both exposures were present but would not occur if just one or the other exposure were present (VanderWeele and Robins, 2007, 2008). If we let D_{ge} denote the counterfactual outcome (the outcome that would have occurred) for each subject if, possibly contrary to fact, G had been set to g and E had been set to e, then a sufficient cause interaction is present if for some individual $D_{11} = 1$ but $D_{10} = D_{01} = 0$. This is in some sense a "mechanistic interaction" insofar as when both exposures are present the outcome is turned "on" but when only one or the other exposure is present the outcome is turned "off". It can furthermore be shown that if such a sufficient cause interaction is present, then within Rothman's sufficient cause framework (Rothman, 1976) there must be a sufficient cause for D which has both G and E as components (VanderWeele and Robins, 2007, 2008). This is thus sometimes called "synergism" between G and E in the sufficient cause framework. Note that a sufficient cause interaction does require some individual with $D_{11} = 1$ but $D_{10} = D_{01} = 0$ but does not require $D_{00} = 0$ for this individual. Further below we will also consider an even stronger notion of "mechanistic interaction" which requires some individual for whom $D_{11} = 1$ and $D_{10} = D_{01} = D_{00} = 0$. However, we will begin our discussion of mechanistic interaction with the slightly weaker notion of a sufficient cause interaction, as this is all that is required for synergism between G and E within the sufficient cause framework.

Additive interaction is sometimes used to test for such mechanistic or sufficient cause interaction. However, having positive additive interaction only implies such sufficient cause interaction under additional assumptions. If it can be assumed that both exposures are never preventive for any individual (formally, if D_{ge} is non-decreasing in g and e for all individuals), then provided control is also made for confounding of both exposures,[10] positive additive interaction, $p_{11} - p_{10} - p_{01} + p_{00} > 0$, suffices for sufficient cause interaction (Greenland et al., 2008; VanderWeele and Robins, 2007). The assumption that neither exposure can ever be preventive for any individual is sometimes referred to as a positive "monotonicity" assumption; it is a strong assumption. In some contexts, it might be plausible. For example, we would probably never think that smoking is protective for lung cancer for any individual. There may be some persons for whom smoking causes lung cancer, there may be others for whom smoking is neutral, but we would never think that smoking prevents lung cancer for anyone (i.e. that they would not have lung cancer if they smoked, but that they would have lung cancer if they did not smoke). Thus the positive monotonicity assumption for the effect of smoking on lung cancer may be plausible. But in other cases the assumption may be less plausible. For example, if we were to consider the effect of alcohol consumption on stroke, alcohol may be protective for stroke in some persons but causative for others; the monotonicity assumption would not be plausible here. Positive monotonicity requires that the effect is never preventive for the outcome for any person in the population. Importantly, to assess sufficient cause interaction simply by examining whether additive interaction is positive requires that the effects of both exposures on the outcome be monotonic. This will in many contexts be a strong assumption, and it is an assumption that is not possible for verify empirically; it must be established on substantive grounds.

Fortunately, it is also possible to test for sufficient cause interaction even without such monotonicity assumptions but the standard tests for positive additive interaction no longer suffice. Alternative tests must

10 Formally, we say that the effects of both exposures are unconfounded if the counterfactual outcomes D_{ge} are independent of the actual exposures $\{G, E\}$; or that the effect of both exposures are unconfounded conditional on covariates C if D_{ge} is independent of exposures $\{G, E\}$ conditional on C.

be used. VanderWeele and Robins (2007, 2008) showed that if the effect of the two exposures were unconfounded then

$$p_{11} - p_{10} - p_{01} > 0$$

would imply the presence of a sufficient cause interaction. This is a stronger condition than regular positive additive interaction which only requires $p_{11} - p_{10} - p_{01} + p_{00} > 0$, because, with the condition $p_{11} - p_{10} - p_{01} > 0$, we are no longer adding back in the outcome probability p_{00} for the doubly unexposed group. This condition for a sufficient cause interaction, without making monotonicity assumptions, thus does not correspond to, and is stronger than, the regular test for additive interaction, or than simply examining whether interaction is positive in a statistical model (VanderWeele, 2009b). In these various cases, the magnitude of the contrast $p_{11} - p_{10} - p_{01} + p_{00}$ with monotonicity or $p_{11} - p_{10} - p_{01}$ without monotonicity in fact gives a lower bound on the prevalence of individuals manifesting sufficient cause interaction patterns (VanderWeele et al., 2010).

If data are only available on the ratio scale, then if both exposures have positive monotonic effects on the outcomes, we can test for sufficient cause interaction by the condition $RERI_{RR} > 0$. Likewise, the condition $p_{11} - p_{10} - p_{01} > 0$ without imposing monotonicity assumptions can be expressed in terms of $RERI_{RR}$ as $RERI_{RR} > 1$; again this is stronger than simply the ordinary condition for additive interaction $RERI_{RR} > 0$. However, $RERI_{RR}$ still can be used in a straightforward way to test for such sufficient cause interaction by testing whether $RERI_{RR} > 1$ rather than simply $RERI_{RR} > 0$.

Note that when the empirical conditions above are satisfied, the conclusion is that there are some individuals for whom $D_{11} = 1$ and $D_{10} = D_{01} = 0$; the conclusion is not that all individuals have this response pattern. Note also that these conditions given here are sufficient but not necessary for sufficient cause interaction, i.e. if these conditions are satisfied then a sufficient cause interaction must be present, but if the conditions are not satisfied, then there may or may not be a sufficient cause interaction − one simply cannot determine this from the data. The conditions given here are the weakest possible empirical conditions to test for sufficient cause interaction without making further assumptions (VanderWeele and Richardson, 2012).

VanderWeele (2010a, 2010b) discussed empirical tests for an even stronger notion of interaction. We might say that there is a "singular" or "epistatic" interaction if there are individuals in the population who will have the outcome if and only if both exposures are present; in counterfactual notation, that is, there are individuals for whom $D_{11} = 1$ but $D_{10} = D_{01} = D_{00} = 0$. In the genetics literature, when gene−gene interactions are considered, such response patterns are sometimes called instances of "compositional epistasis" (Phillips, 2008; Cordell, 2009) and constitute settings in which the effect of one genetic factor is masked unless the other is present. VanderWeele (2010a, 2010b) noted that if the effects of the two exposures on the outcome were unconfounded then

$$p_{11} - p_{10} - p_{01} - p_{00} > 0$$

would imply the presence of such an "epistatic interaction". Again this is an even stronger notion of interaction; in this condition for "epistatic interaction" we are now subtracting p_{00}. The condition $p_{11} - p_{10} - p_{01} - p_{00} > 0$ expressed in terms of $RERI_{RR}$ is equivalent to $RERI_{RR} > 2$.

For epistatic interactions, if the effect of at least one of the exposures is positive monotonic (Y_{ge} is non-decreasing in at least one of g or e), then $p_{11} - p_{10} - p_{01} > 0$ suffices for an epistatic interaction and tests for $RERI_{RR} > 1$ could be used; if the effects of both exposures are positive and monotonic, then $p_{11} - p_{10} - p_{01} + p_{00} > 0$ suffices and tests for $RERI_{RR} > 0$ could be used to test for an epistatic interaction (VanderWeele, 2010a, 2010b). These conditions are likewise sufficient but not necessary for an epistatic interaction; if these conditions are satisfied, then an epistatic interaction must be present, but if the conditions are not satisfied, then an epistatic interaction may or may not be present. Note also that when the empirical conditions above are satisfied, the conclusion is that there are some individuals for whom $D_{11} = 1$ but $D_{10} = D_{01} = D_{00} = 0$; the conclusion is not that all individuals have this response pattern. The various results are summarized in Table 7.

Table 7 Relations between the additive relative excess risk due to interaction (*RERI*) and forms of mechanistic interaction under different monotonicity assumptions ("S" indicates the presence of a sufficient cause interaction; "E" denotes an epistatic interaction)

Monotonicity assumption	$RERI_{RR} > 0$	$RERI_{RR} > 1$	$RERI_{RR} > 2$
No assumptions about monotonicity	–	S	S,E
One of *G* or *E* have positive monotonic effects	–	S,E	S,E
Both *G* and *E* have positive monotonic effects	S,E	S,E	S,E

The sufficient conditions given here for mechanistic interaction require that control has been made for confounding of the effects of both exposures. The sufficient conditions for $RERI_{RR}$ for mechanistic interaction in Table 7 still apply when adjustment is made for confounders, e.g. when the relative excess risk due to interaction is calculated using logistic regression as described above adjusting for covariates.[11]

In assessing additive interaction using $RERI_{RR}$, it thus is useful to examine not only whether the estimate and confidence interval for $RERI_{RR}$ are greater than 0 (i.e. whether there is additive interaction) but also whether the estimate and confidence interval for $RERI_{RR}$ are all greater than 1 or are all greater than 2. This is because $RERI_{RR}$ of this magnitude would provide evidence for mechanistic interaction (sufficient cause or epistatic interaction) without the need for additional assumptions. The $RERI_{RR}$ scale is in some sense the natural scale on which to assess mechanistic interaction and has the thresholds of 0, 1, and 2 for varying degrees of evidence (according to the strength of the assumptions needed for the conclusion). We noted above that $RERI_{RR}$ cannot be used to assess the magnitude of the underlying additive interaction for risks, but we see here that although its magnitude does not necessarily correspond to the magnitude of the additive interaction for risks, the magnitude of $RERI_{RR}$ does give differing degrees of evidence for mechanistic interaction.[12]

As an example, Bhavnani et al. (2012), using age-standardized measures, reported that risk ratios for diarrheal disease across groups infected with rotavirus and/or Giardia. With the doubly unexposed group as the reference category, the risk ratio for rotavirus (in the absence of Giardia) is 2.63, the risk ratio for Giardia (in the absence of rotavirus) is 1.13, and the risk ratio when both rotavirus and Giardia

11 When statistical models are used to adjust for confounding, this requires correct model specification. Within Rothman's sufficient cause framework, such statistical models can impose constraints on the sufficient causes which are sometimes thought undesirable (VanderWeele et al., 2010). In such cases, alternative modeling approaches using weighting or semiparametric methods can help relax these modeling assumptions (Vansteelandt et al., 2008, 2012; VanderWeele et al., 2010; VanderWeele and Vansteelandt, 2011) but are beyond the scope of the current tutorial.

12 Testing for sufficient cause or epistatic interaction can also be done simply by using the interaction parameter of a log-linear model (or logistic model if odds ratios approximate risk ratios) directly. The log-linear model for risk ratios that includes a product term takes the form: $\log\{P(Y=1|G=g,E=e)\} = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 eg$. Here, if both *G* and *E* have positive monotonic effects on *Y*, then the condition $\beta_3 > 0$ implies both a sufficient cause interaction and an epistatic interaction (VanderWeele, 2009b, 2010b). If at least one of *G* or *E* have positive monotonic effects on *Y*, then, provided both the main effects of *G* and *E* on *Y* are non-negative (i.e. $\beta_1 \geq 0$ and $\beta_2 \geq 0$), the condition $\beta_3 > \log(2)$ implies both a sufficient cause interaction and an epistatic interaction (VanderWeele, 2009b, 2010b). Since $e^{\beta_3} = RR_{11}/(RR_{10}RR_{01})$ this is just equivalent to the condition for the multiplicative risk ratio interaction $RR_{11}/(RR_{10}RR_{01}) > 2$. If neither *G* nor *E* has positive monotonic effects on *Y*, then, provided both the main effects of *G* and *E* on *Y* are non-negative (i.e. $\beta_1 \geq 0$ and $\beta_2 \geq 0$), the condition $\beta_3 > \log(2)$ implies a sufficient cause interaction and the condition $\beta_3 > \log(3)$ implies an epistatic interaction (VanderWeele, 2009b, 2010b). Thus, once again, without monotonicity assumptions a positive statistical multiplicative interaction, $\beta_3 > 0$, alone does not suffice and we need stronger conditions e.g. $\beta_3 > \log(2)$ or $\beta_3 > \log(3)$. However, if we can estimate the parameters of the multiplicative model $\beta_1, \beta_2, \beta_3$ then, as described above, we can calculate the relative excess risk due to interaction by $RERI_{RR} = e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} - e^{\beta_2} + 1$ and we would be better off testing for sufficient cause synerigsm using the conditions $RERI_{RR} > 0$ or $RERI_{RR} > 1$ or $RERI_{RR} > 2$, respectively, as these conditions are more often satisfied than those for the multiplicative interaction ($\beta_3 > 0$, $\beta_3 > \log(2)$, and $\beta_3 > \log(2)$); the multiplicative interaction conditions imply the relative excess risk due to interaction conditions, but not vice versa. The comments here for statistical interaction for risk ratios in a log-linear model pertain also approximately to statistical interaction for odds ratios in a logistic regression model when the outcome is rare.

are present is 10.72. This gives $RERI_{RR} = 10.72 - 2.63 - 1.13 + 1 = 7.96$ (95% CI: 3.13, 18.92). The value of $RERI_{RR}$ and its entire 95% confidence interval exceed the value 2, suggesting strong evidence for mechanistic interaction (both "sufficient cause" and "epistatic" interaction) even in the absence of any monotonicity assumptions.

Although it is beyond the scope of the current paper extensions of these ideas are available for exposures with more than two levels (VanderWeele, 2010a, 2010b, 2010d) and for multi-way interactions between three or more exposures (VanderWeele and Robins, 2008; VanderWeele and Richardson, 2012) as well as for settings with causal antagonism in which the presence of one exposure may block the operation of the other (VanderWeele and Knol, 2011b) See VanderWeele (2014b) for an overview. In the next section, we will discuss how even these so-called mechanistic interactions (sufficient cause or epistatic interactions) considered here give limited information about the underlying biology.

2 Part II: Limitations, extensions, study design, and properties of interaction analysis

2.1 Limits of inference concerning biology

Although tests for sufficient cause interaction, like those considered in the previous section, can shed light on whether there are individuals for whom the outcome would occur if both exposures are present but not if just one or the other is present, it should be noted, that even such "mechanistic interaction", does not imply that the two exposures are *physically* interacting in any real sense (Siemiatycki and Thomas, 1981; Thompson, 1991; VanderWeele and Robins, 2007; Phillips, 2008; Cordell, 2009). To see this, suppose that G_1 and G_2 are two genetic factors. Suppose that when $G_1 = 1$ protein 1 is not produced and that when $G_2 = 1$ protein 2 is not produced. Suppose that the outcome D occurs if and only if neither protein 1 nor protein 2 is present. We then have an epistatic interaction because the outcome occurs if and only if $G_1 = 1$ and $G_2 = 1$, but we do not have physical interaction here. It is precisely the absence of the proteins that gives rise to the outcome; there simply is nothing to physically interact here.

We should thus distinguish between (i) statistical interaction on the one hand and (ii) mechanistic interaction (e.g. the outcome occurs if both exposures are present but not if just one or the other is present) on the other, and finally, (iii) "biological" or "functional" interaction in which the two exposures physically interact to bring about the outcome (Phillips, 2008; Cordell, 2009; VanderWeele, 2010a, 2011a). In the example just given, we have mechanistic interaction but not "functional" or physical interaction. Thus, although we can sometimes empirically draw conclusions about mechanistic interaction from data, empirical tests will not in general allow us to draw conclusions about functional or physical interaction between exposures and it is important to understand the limits of the conclusions being drawn about these alternative forms of interaction.

Other examples of the limitation of biologic inference concerning interaction were given by Siemiatycki and Thomas (1981). Consider, for example, a setting in which for the outcome to occur two stages of disease development must take place. Several theories for the development of cancer follow this model. Suppose that the two exposures of interest, G_1 and G_2 say, affect different stages: G_1 acts on stage 1 and G_2 acts on stage 2. Suppose also in this example that stage 1 and stage 2 are completely independent of each other. Assume that the baseline probability of stage 1 occurring is 1% and the baseline rate of stage 2 occurring is also 1%, so that the baseline likelihood of disease is 0.01%. Suppose that G_1 increases the probability of stage 1 occurring from 1% to 2% and G_2 increases the probability of stage 2 occurring from 1% to 5%. Suppose, however, that the presence of G_2 in no way alters the effect of G_1's increasing the probability of stage 1 occurring from 1% to 2%; i.e. the probability of stage 2 is 1% if $G_1 = 0$ and 2% if $G_1 = 2$, irrespective of whether G_2 is present or absent. Suppose, similarly, that the presence of G_1 in no way alters the effect of

G_2's increasing the probability of stage 2 occurring from 1% to 5%. Here then we seem to have no interaction between G_1 and G_2 at the biologic level.

As noted above, if neither exposure is present ($G_1 = 0$ and $G_2 = 0$), then the risk of stage 1 and stage 2 are both 1% and the overall likelihood of the outcome is $1\% \times 1\% = 0.01\%$. If just G_1 is present ($G_1 = 1$ and $G_2 = 0$), then the risk of stage 1 is 2% and the risk of stage 2 is 1% and the overall likelihood of the outcome is $2\% \times 1\% = 0.02\%$. If $G_1 = 0$ and $G_2 = 1$, then the risk of stage 1 is 1% and the risk of stage 2 is 5% and the overall likelihood of the outcome is $1\% \times 5\% = 0.05\%$. If $G_1 = 1$ and $G_2 = 1$, then the risk of stage 1 is 2% and the risk of stage 2 is 5% and the overall likelihood of the outcome is $2\% \times 5\% = 0.10\%$. In this example, our measure of multiplicative interaction is $\frac{p_{11}p_{00}}{p_{10}p_{01}} = \frac{0.10\%(0.01\%)}{0.02\%(0.05\%)} = 1$. However, our measure of additive interaction is

$$p_{11} - p_{10} - p_{01} + p_{00} = 0.10\% - 0.02\% - 0.05\% + 0.01\% = 0.04\% > 0.$$

We have positive additive interaction but no biologic interaction in this example. Here our conditions for sufficient cause interaction are satisfied, since

$$p_{11} - p_{10} - p_{01} = 0.10\% - 0.02\% - 0.05\% = 0.03\% > 0,$$

and even our conditions for "epistatic" or "singular" interaction,

$$p_{11} - p_{10} - p_{01} - p_{00} = 0.10\% - 0.02\% - 0.05\% - 0.01\% = 0.02\% > 0,$$

are also satisfied. But again we saw that there was no interaction between G_1 and G_2 at the biologic level. How are we to make sense of this? What we can conclude from the condition for a epistatic or singular interaction, say, are that there are some individuals who would have the outcome if both exposures were present but who would not if just one or the other or neither exposure were present. But we see here that not even this necessarily indicates interaction at some fundamental biologic level. We have this form of "singular" or "sufficient cause" interaction because, if both exposures are present, 0.10% have the outcome and this cannot be accounted by those individuals whose outcome only required the first exposure (0.02%) or only the second (0.05%) or who required neither (0.01%). Even if these three groups were mutually exclusive, they would not account for the risk of 0.10% that occurs if both exposures are present ($0.10\% - (0.02\% + 0.05\% + 0.01\%) = 0.02\% > 0$). There must be some individuals for whom the outcome occurs if and only if both exposures are present. But again, this does not, as this example shows, indicate biologic interaction in any fundamental biologic sense.[13]

We can assess statistical interaction (on any scale we choose), we can assess additive interaction to determine how best to allocate interventions, and we can assess "sufficient cause" or "epistatic/singular" interaction to determine whether there are individuals who would have the outcome if both exposures were present but not if only one or the other were present. All of these may provide some insight into the underlying biology, but we have no way of going from any of these forms of interaction which we can assess with data directly to the underlying biology itself.

13 On the basis of these and other similar examples, Thompson (1991) suggested that if an outcome required stages and one exposure affected the first stage and another exposure affected the second stage (a "multi-stage model"), then if there were no biologic interaction, we would expect a multiplicative model. Likewise he suggested that if the occurrence of a single adverse event was sufficient for the development of the disease (a "single-hit model") then the absence of biologic interaction we would expect an additive model. Finally, he suggested that if the outcome occurred if an individual failed to experience any of one or more occurrences of a beneficial event (a "no-hit model", cf. Walter and Holford, 1978), then the model should again be multiplicative. While such heuristics may be of some use, if we do find that an additive model fits well it is not necessarily the case that we have a "single-hit model" with no biologic interaction; it could equally be the case that we have a "multi-stage model" in which the factors operate antagonistically. Or if we were to find that the multiplicative model fit well, this does not necessarily indicate a "multi-stage model" with no biologic interaction, but could also be a "single-hit model" in which there was biologic interaction. We cannot in general draw conclusions about the type of biologic model and the presence or absence of biologic interaction simply from the statistical models we use. If we find positive multiplicative interaction, this could be a "multi-stage model" or a "no-hit" model with biologic interaction, or it could be a "single-hit model" with biologic interaction, or it could be a more complicated model with no biologic interaction whatsoever. We cannot tell from the data alone. Our inferences about biology are limited.

In some earlier literature, sufficient cause synergism was sometimes earlier referred to as "biologic interaction" (e.g. Rothman and Greenland, 1998); sometimes even just additive interaction was even referred to as "biologic interaction" (e.g. Andersson et al., 2005). However, as we have seen in the examples above, neither statistical additive interaction nor even sufficient cause interaction or epistatic interaction necessarily tells us anything about physical or functional interactions. Statistical analyses can only tell us limited information about the underlying biology (Siemiatycki and Thomas, 1981; Thompson, 1991; Rothman and Greenland, 1998; Cordell, 2002). Because of this there has been a suggestion to move away from the use "biologic interaction" for sufficient cause interaction or synergism in the sufficient cause framework (cf. Lawlor, 2011; VanderWeele, 2011a). It may be more appropriate to refer to these sufficient cause or epistatic interactions as "mechanistic interactions"; these are still cases in which both exposures together turn the outcome "on" and the removal of one turns the outcome "off" and thus the "mechanistic" description seems potentially appropriate. If even this is thought to be language that is too strong (if "mechanistic" is still thought to indicate biology rather than indicating "on" and "off"), then simply using the terms "sufficient cause interaction" or "singular interaction" may be best.

2.2 Attributing effects to interactions

2.2.1 Attributing joint effects to interactions

At the beginning of the tutorial, we discussed different measures concerning the proportion of risk or effect attributable to interaction. In fact, we can actually decompose the joint effects of the two exposures, G and E, into three components: (i) the effect due to G alone, (ii) the effect due to E alone, and (iii) the effect due to their interaction. On the risk difference scale this decomposition is

$$p_{11} - p_{00} = (p_{10} - p_{00}) + (p_{01} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00}).$$

where the first component, $(p_{10} - p_{00})$, is the effect due to G alone, the second component, $(p_{01} - p_{00})$, is the effect due to E alone, and the final component, $(p_{11} - p_{10} - p_{01} + p_{00})$, is just the standard additive interaction. We could then also compute the proportion of the joint effect due to G alone, $\frac{(p_{10}-p_{00})}{(p_{11}-p_{00})}$, due to E alone, $\frac{(p_{01}-p_{00})}{(p_{11}-p_{00})}$, and due to their interaction, $\frac{(p_{11}-p_{10}-p_{01}+p_{00})}{(p_{11}-p_{00})}$.

We can also carry out a similar decomposition on the ratio scale using excess relative risks. We can decompose the excess relative risk for both exposures, $RR_{11} - 1$, into the excess relative risk for G alone, for E alone, and the excess relative risk due to interaction, $RERI$. Specifically we have (VanderWeele and Tchetgen Tchetgen, 2014)

$$RR_{11} - 1 = (RR_{10} - 1) + (RR_{01} - 1) + RERI_{RR}.$$

We could then likewise compute the proportion of the effect due to G alone, $\frac{RR_{10}-1}{RR_{11}-1}$, due to E alone, $\frac{RR_{01}-1}{RR_{11}-1}$, and due to their interaction $\frac{RERI_{RR}}{RR_{11}-1}$.[14]

14 As discussed at the beginning of the tutorial, Rothman (1986) considered a measure of interaction that he called the attributable proportion, defined as $\frac{RERI}{RR_{11}}$; the denominator Rothman used was RR_{11}. The measure was meant to capture the proportion of the *disease* in the doubly exposed group that is due to the interaction. Rothman (1986) also considered an alternative measure, $\frac{RERI}{RR_{11}-1}$, which captured the proportion of the *effect* of both exposures on the additive scale that is due to interaction. This latter definition is the measure used in the decomposition here (VanderWeele and Tchetgen Tchetgen, 2014). Most of the subsequent literature has focused on the former measure; but the latter measure, i.e. using $RR_{11} - 1$, as the denominator in fact has some advantages (VanderWeele, 2013). With Rothman's primary measure, $\frac{RERI}{RR_{11}}$, even if all of the joint effect were due to interaction so that the effect of G alone and E alone were both risk ratios of 1, i.e. $RR_{10} = 1$ and $RR_{01} = 1$, we would nevertheless have that Rothman's primary attributable proportion measure would be $\frac{RERI}{RR_{11}} = \frac{RR_{11}-RR_{10}-RR_{01}+1}{RR_{11}} = \frac{RR_{11}-1-1+1}{RR_{11}} = \frac{RR_{11}-1}{RR_{11}} < 1$ i.e. even if the entirety of the joint effect of both exposures were due to interaction, the attributable proportion measure is still less than 100%. The measure $\frac{RERI}{RR_{11}-1}$ does not have this issue. It is 100% when the

Under the logistic regression model

$$\text{logit}\{P(D=1|G=g,E=e,C=c)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg + \gamma_4' c. \quad [9]$$

for an outcome that is rare, the joint effect attributable to G alone, E alone, and to their interaction are given approximately by:

$$\frac{RR_{10}-1}{RR_{11}-1} \approx \frac{e^{\gamma_1}-1}{e^{\gamma_1+\gamma_2+\gamma_3}-1},$$

$$\frac{RR_{01}-1}{RR_{11}-1} \approx \frac{e^{\gamma_2}-1}{e^{\gamma_1+\gamma_2+\gamma_3}-1},$$

$$\frac{RERI_{RR}}{RR_{11}-1} \approx \frac{(e^{\gamma_1+\gamma_2+\gamma_3}-e^{\gamma_1}-e^{\gamma_2}+1)}{e^{\gamma_1+\gamma_2+\gamma_3}-1}.$$

The expressions can be used even when control is made for covariates in the logistic regression. VanderWeele and Tchetgen Tchetgen (2014) provided SAS and Stata code to do this automatically and to calculate standard errors and confidence intervals for the proportions and also discussed extensions to exposures that are not binary. Note that to interpret the effects above causally, one would have to control for confounding of the relationships of both exposures with the outcome.

We illustrate the various decompositions with an example from genetic epidemiology presented by VanderWeele and Tchetgen Tchetgen (2014) using data from a case–control study of lung cancer at Massachusetts General Hospital of 1,836 cases and 1,452 controls (Miller et al., 2002). The study included information on smoking and genotype information on locus 15q25.1. For simplicity, we will code the exposure as binary so that smoking is ever versus never and the genetic variant is a comparison of 0 versus 1/2 T alleles at rs8034191. Analyses were restricted to Caucasians, and covariate data include age (continuous), gender, and educational history (college degree or more, yes/no). If we proceed with the decomposition of the joint effect, then the proportions attributable to G alone, E alone, and to their interaction are

$$\frac{RR_{10}-1}{RR_{11}-1} \approx 0.8\%(95\%\text{CI}: -6.2\%, 7.7\%),$$

$$\frac{RR_{01}-1}{RR_{11}-1} \approx 51.4\%(95\%\text{CI}: 33.4\%, 69.4\%),$$

$$\frac{RERI}{RR_{11}-1} \approx 47.8\%(95\%\text{CI}: 33.3\%, 62.3\%).$$

main effects of G alone and E alone were both risk ratios of 1 i.e. when the entirety of the joint effect is due to interaction. The measure $\frac{RERI}{RR_{11}-1}$ captures the proportion of the joint effect attributable to interaction. The attributable proportion of joint effects measure, $\frac{RERI}{RR_{11}-1}$, is also attractive from another standpoint. Skrondal (2003) criticized Rothman's original attributable proportion measure because, in the presence of covariates, if the risks follow a linear risk model that is additive in the covariates, $P(D=1|G=g,E=e,C=c) = \alpha_0 + \alpha_1 g + \alpha_2 e + \alpha_3 ge + \alpha_4 c$, then, although the additive interaction, $p_{11} - p_{10} - p_{01} + p_{00} = \alpha_3$, does not vary across strata of the covariates, Rothman's primary attributable proportion measure, $\frac{RERI}{RR_{11}} = \frac{\alpha_3}{\alpha_0+\alpha_1+\alpha_2+\alpha_3+\alpha_4 c}$, does vary across strata of the covariates. Skrondal also noted that *RERI* itself, which would be given here by $RERI = \frac{\alpha_3}{\alpha_0+\alpha_4 c}$, likewise depends on the covariates. However, the measure of the proportion of the joint effects attributable to interaction, $\frac{RERI}{RR_{11}-1} = \frac{\alpha_3}{\alpha_1+\alpha_2+\alpha_3}$, does not vary with the covariates and thus circumvents Skrondal's criticism. Likewise, the other two components in the decomposition: $\frac{RR_{10}-1}{RR_{11}-1} = \frac{\alpha_1}{\alpha_1+\alpha_2+\alpha_3}$ and $\frac{RR_{01}-1}{RR_{11}-1} = \frac{\alpha_2}{\alpha_1+\alpha_2+\alpha_3}$ also do not depend on the covariates. The decomposition of the joint effect of the two exposures into three components, (i) the effect due to G alone, (ii) the effect due to E alone, and (iii) the effect due to their interaction, thus entirely circumvents Skrondal's critique of *RERI* and Rothman's primary attributable proportion measure, $\frac{RERI}{RR_{11}}$.

Almost none of the joint effect (comparing both G and E present to both absent) is due to the effect of G in the absence of E, about 51% is due to E is the absence of G and about 48% is due to the interaction between G and E.

2.2.2 Attributing total effects to interactions

If the distribution of two exposures G and E are independent (i.e. uncorrelated) in the population, then we can also decompose the total effect of one of the exposures (e.g. total effect of E) into two components (VanderWeele and Tchetgen Tchetgen, 2014). If we let p_e denote $P(D=1|E=e)$, i.e. the probability that $D=1$ when $E=e$ then we have

$$(p_{e=1} - p_{e=0}) = (p_{01} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(G=1).$$

This decomposes the overall effect of E on Y into two pieces: the first piece is the conditional effect of E on Y when $G=0$, the second piece is the standard additive interaction, $(p_{11} - p_{10} - p_{01} + p_{00})$, multiplied by the probability that $P(G=1)$. In some sense then we can attribute the total effect of E on Y to the part that would be present still if G were 0 (this is $p_{01} - p_{00}$) and to a part that has to do with the interaction between G and E (this is $(p_{11} - p_{10} - p_{01} + p_{00})P(G=1)$). If we could remove the genetic exposure, i.e. set it to 0, we would remove the part that is due to the interaction and we be left with only $p_{01} - p_{00}$. Since we can do this decomposition we might define a quantity $pAI_{G=0}(E)$ as the proportion of the overall effect of E that is attributable to interaction, with a reference category for the genetic exposure of $G=0$, as

$$pAI_{G=0}(E) := \frac{(p_{11} - p_{10} - p_{01} + p_{00})P(G=1)}{(p_{e=1} - p_{e=0})}.$$

The remaining portion $(p_{01} - p_{00})/(p_{e=1} - p_{e=0})$ is the proportion of the effect of E that would remain if G were fixed to 0. VanderWeele and Tchetgen Tchetgen (2014) provided SAS and Stata code to do this automatically and handle more general cases and models. Note that the three-way decomposition above for *joint* effects did not require that the exposures be independent of one another. However, the two-way decomposition for a *total* effect given here in general assumes that the exposures are independent. VanderWeele and Tchetgen Tchetgen (2014) and VanderWeele (2014a) also discussed similar, but more complex, decompositions when the two exposures, G and E, are correlated.

As already discussed in this tutorial, one of the motivations for studying interaction is to identify which subgroups would benefit most from intervention when resources are limited. In settings in which it is not possible to intervene directly on the primary exposure of interest, one might instead be interested in which other covariates could be intervened upon to eliminate much or most of the effect of the primary exposure of interest. The methods here for attributing effects to interactions can be useful in assessing this and identifying the most relevant covariates for intervention.

2.3 Case-only designs

Another more recent approach concerning statistical interaction is also worth noting. Consider the statistical interaction β_3 in the log-linear model:

$$\log\{P(D=1|G=g, E=e)\} = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 eg.$$

Suppose now also that the distribution of the two exposures, G and E, are independent in the population. This assumption may be plausible in many gene−environment interaction studies. Suppose further that data are only collected on the cases ($D=1$). It can be shown that under this independence assumption, the odds ratio relating G and E among the cases is equal to the interaction measure on the multiplicative scale β_3 (Yang et al., 1999; cf. Piergorsch et al., 1994):

$$\frac{P(G=1|E=1,D=1)/P(G=0|E=1,D=1)}{P(G=1|E=0,D=1)/P(G=0|E=0,D=1)} = \frac{RR_{11}}{RR_{10}RR_{01}} = \beta_3.$$

Somewhat surprisingly, to get measures of multiplicative interaction, all that is needed is data on G and E among the cases. The use of the odds ratio relating G and E among the cases is referred to as the "case-only" estimator of interaction. With the case-only estimator we can estimate the interaction parameter β_3, but we cannot estimate the main effects of the log-linear regression, β_1 and β_2.

The case-only estimator depends critically on the assumption that the distribution of the two exposure are independent in the population and can be quite biased if this assumption is violated (Albert et al., 2001). However, under this assumption of independence in distribution, the case-only estimator is in fact more efficient than using the standard estimate from a log-linear regression (Yang et al., 1997).

The same result holds for statistical interaction in logistic regression

$$\text{logit}\{P(D=1|G=g,E=e)\} = \gamma_0 + \gamma_1 g + \gamma_2 e + \gamma_3 eg$$

under the assumption that the outcome is rare (Piergorsch et al., 1994). The result for log-linear models does not require a rare outcome. Sometimes, for logistic regression, the independence assumption is articulated as one of independence of G and E among the non-cases. For a rare outcome, this is approximately equivalent to independence in the population.

The result also holds for log-linear or logistic regression if we control for covariates. The conditional independence assumption is then that the distributions of G and E are independent conditional on C. Estimates and confidence intervals for the case-only estimator can be obtained by running a logistic regression of G on E and C among the cases:

$$\text{logit}\{P(G=1|E=e,C=c,D=1)\} = \theta_0 + \theta_1 e + \theta_2' c.$$

The coefficient and confidence interval for θ_1 in this regression on the cases will equal that of the product term coefficient in the log-linear model with covariates provided the distributions of G and E are independent in the population and will equal the product term coefficient in the logistic model with covariates, in addition, that the outcome is rare.

Note that in all of these cases, to interpret the multiplicative interaction parameter estimate from the statistical model as causal interaction on a multiplicative scale, it would be necessary to assume that the effects of both exposures on the outcome are unconfounded (conditional on covariates C). To interpret the parameter estimate as a measure of effect heterogeneity on the multiplicative scale, it would be necessary to assume that the effect of one of the exposures on the outcome is unconfounded (conditional on covariates C). In a case-only study, simply assuming that the effect of one exposure on the other exposure is unconfounded does not suffice to give a causal interpretation for the effects of either or both exposures on the outcome Y.

As an example, Bennet et al. (1999) used data on non-smoking lung cancer cases and reported exposure status for GSTM1 genotype and passive smoking as in Table 8.

Table 8 Number of cases by genotype and smoking status (Bennett et al., 1999)

	No smoking	Smoking
GSTM1 present, $G=0$	28	14
GSTM1 absent, $G=1$	27	37

Using data only on the cases we have that the estimate of multiplicative interaction is

$$\frac{RR_{11}}{RR_{10}RR_{01}} = \frac{P(G=1|E=1,D=1)/P(G=0|E=1,D=1)}{P(G=1|E=0,D=1)/P(G=0|E=0,D=1)} = \frac{37/14}{27/28} = 2.74.$$

When adjusted also for age, radon exposure, saturated fat intake, and vegetable intake using logistic regression, the case-only estimate of multiplicative interaction is 2.6 (95% CI: 1.1−6.1). There is evidence here for multiplicative interaction between passive smoking and the absence of GSTM1 on lung cancer.

VanderWeele et al. (2010) discussed using the case-only estimator to assess mechanistic interaction and showed that if the main effects of both exposures are non-negative (which cannot be assessed directly in a case-only study but could be evaluated on substantive grounds), then a sufficient cause interaction is present if $\theta_1 > \log(2)$ without any individual level monotonicity assumptions, or if $\theta_1 > 0$ when it can be assumed that both exposures have positive monotonic effects on the outcome. They also noted that if the main effects of both exposures are non-negative then an epistatic interaction is present if $\theta_1 > \log(3)$ without any individual level monotonicity assumptions, or if $\theta_1 > \log(2)$ and at least one of the two exposures has a positive monotonic effect, or if $\theta_1 > 0$ and both exposures have positive monotonic effects.

2.4 Interactions for continuous outcomes

When continuous outcomes are in view, linear and log-linear regression can still be used to estimate measures of additive and multiplicative interaction, respectively. For additive interaction, a linear regression model for the continuous outcomes could be used

$$E(D|G = g, E = e, C = c) = \alpha_0 + \alpha_1 g + \alpha_2 e + \alpha_3 eg + \alpha_4' c,$$

and α_3 can be taken as a measure of additive interaction. This parameter is equal to the additive interaction measure:

$$\alpha_3 = E(D|G = 1, E = 1, C = c) - E(D|G = 1, E = 0, C = c)$$

$$- E(D|G = 0, E = 1, C = c) + E(D|G = 0, E = 0, C = c).$$

For multiplicative interaction, a log-linear regression model for the continuous outcomes could be used

$$\log\{E(D|G = g, E = e, C = c)\} = \beta_0 + \beta_1 g + \beta_2 e + \beta_3 eg + \beta_4' c,$$

and β_3 can be taken as a measure of multiplicative interaction. This parameter, when exponentiated, is equal to the multiplicative interaction measure:

$$e^{\beta_3} = \frac{E(D|G = 1, E = 1, C = c)/E(D|G = 1, E = 0, C = c)}{E(D|G = 0, E = 1, C = c)/E(D|G = 0, E = 0, C = c)}.$$

Note that with a continuous outcome most of the arguments for preferring one scale to another are no longer applicable. With a continuous outcome, we generally no longer run into convergence problems for the additive scale. But the argument for the public health significance of the additive scale is not as applicable for a continuous outcome as we are no longer analyzing discrete events. Moreover, with a continuous outcome, it is not clear that the additive scale gives any insight into mechanistic interaction. Whether additive or multiplicative scales are to be preferred for a continuous outcome will generally depend on the distribution of the outcome data.

2.5 Identifying subgroups to target treatment using multiple covariates

Thus far our focus has been on estimating and interpreting interactions; we have focused on binary exposures but have also briefly considered ordinal or continuous exposures. As we had noted above, one motivation for examining interaction is determining whether a particular intervention might be more effective for one subgroup than another. It was noted that assessing interaction on the additive scale was most important for this purpose. This motivation does, however, raise the question as to how to choose the variable or

variables that are to define subgroups. Most of our discussion has presupposed that we have a particular secondary variable in mind which will define subgroups and for which we will examine whether there is effect heterogeneity across subgroups. In some settings, data on many such variables that could potentially define subgroups may be available. One option would then be to use each of these and see if any of them are such that there is evidence for substantial effect heterogeneity. A downside of this approach is that by testing for effect heterogeneity across many variables, we are more likely to find spurious results suggesting effect heterogeneity by chance. We would need to correct for such "multiple testing" to mitigate this possibility, and this is often done by using a Bonferroni correction in which the *p*-value cutoff (typically 0.05) is divided by the number of tests conducted to give a more stringent threshold. An alternative approach and one that is often advocated in the literature is to decide in advance, based on substantive knowledge, which factor or factors are thought most likely to show evidence for effect heterogeneity and test for these alone.

An additional complication arises when the variable that is going to define subgroups is continuous. One might then have to decide what cutoff of the continuous variable is to be used in defining subgroups. One might also be interested in whether there is in some sense an optimal cutoff of such a continuous variable such that whenever the variable is above that level it is best to treat. Methods to address this type of question are now available for a single continuous variable (Bonetti and Gelber, 2000, 2005; Song and Pepe, 2004).

However, further complications arise when one is interested in using multiple continuous or categorical variables simultaneously. An even more general approach involves forming anticipated "effect scores" for each and every person in a sample or population based on many baseline covariates and then targeting treatment to those above a certain "effect score" threshold. One approach to forming such effect scores is to fit a regression model for the outcome on all or several covariates for the treated or exposed subjects and then to fit a separate model for the untreated or unexposed subjects. For each person in the sample one can then use the two models, once they are fit to the data, to get a predicted outcome (or probability of the outcome) under exposure and a predicted outcome (or probability of the outcome) under control. The difference between these two predicted outcomes would then be the individual "effect score." One might then consider targeting treatment to those only above a certain threshold. This approach has the advantage of being able to incorporate information from many different covariates in defining subgroups to try to optimize the effect of treatment. It would even be possible to compare different models for the outcome under the exposed and control conditions, or different sets of covariates, in these models, to see which has the "effect scores" that best allows one to predict the outcome and target subpopulations (Zhao et al., 2013).

The approach is appealing and intuitive. Several complications do, however, arise in trying to make inferences in this manner, though methods have been developing to help address these. One complication is "overfitting": if the same data are used to fit the models and to evaluate which of the effect scores, and models, and covariates, have the best predictive properties in forming subgroups, then the performance in a different sample might not be very good. Because of the potential for overfitting, the evaluation of the effect scores and models and covariates may be misleading because the model parameters were specifically estimated to fit the available data as best as possible, and if the same parameters were used to get predicted outcomes in a different sample drawn from the same population, its performance would not be as good. Zhao et al. (2013) have proposed a cross-validation procedure which involves splitting the sample into a training dataset (which is used to fit the models) and an evaluation dataset (which is used to evaluate and compare effects scores and models and covariates) to address this problem. Based on simulations they recommend using 4/5 of the data to fit the models and 1/5 to evaluate the models.

Another complication that can arise with this effect-score approach is that if the models to get predicted outcomes are not correctly specified then the inferences about the effects for different subgroups defined by the effect score may be misleading. Cai et al. (2011) have proposed a two-stage approach which helps address this issue. They recommend fitting parametric regression model for the treated and control subjects to form the effect scores and then to use non-parametric regression to estimate the effects of the treatment on the outcome across subgroups defined by these effect scores. They describe procedures to carry out inference and form confidence intervals for the effects across subgroups defined by the effect scores that are applicable even if the parametric models initially used to form the effect scores are not correctly specified.

These approaches using multiple covariates to identify subgroups for which to target treatment are appealing and potentially powerful. More methodological development remains to be done so that these are easy to implement and optimally choose cutoffs but as these methods develop it is likely they will be very useful in both observational and experimental research.

2.6 Robustness of interaction to unmeasured confounding and sensitivity analysis

As noted earlier, if we are interested in estimates of causal interaction, e.g. assessing what the effects on the outcome would be if we were to intervene on both exposures, then we have to control for confounding for both the exposures. If we have failed to control for confounding, then our interaction estimates may be biased. There are, however, cases in which unmeasured confounding will not bias estimates of interaction. Specifically suppose we had an unmeasured confounder U of one of the exposures, say E, then if the distributions of G and E are independent in the population, and if U does not interact with G on the additive scale then estimates of additive interaction will be unbiased even if control is not made for U (VanderWeele et al., 2012) and even though the main effect for E is thus biased. Likewise, if G and E are independent, and if U does not interact with G on the multiplicative scale then estimates of multiplicative interaction will be unbiased even if control is not made for U (VanderWeele et al., 2012). Analogous results hold if the unmeasured confounder affects G rather than E, and analogous results also hold in some cases in which there are unmeasured confounders of G and of E (VanderWeele et al., 2012); the independence assumption can also be somewhat relaxed (Tchetgen Tchetgen and VanderWeele, 2012). Finally, if these assumptions of independence and no interaction between U and G or E fail, then sensitivity analysis techniques for interaction on the additive or multiplicative scale (VanderWeele et al., 2012) can be employed to assess how robust one's conclusions about interaction are to unmeasured confounding. Note also that, as discussed above, if only one of the two exposures is subject to confounding then (even without controlling for such confounding), interaction estimates can sometimes still be interpreted as measures of effect heterogeneity (i.e. how interventions on the effect of one exposure vary across strata defined by the second exposure, where we do not intervene on the second exposure).

2.7 Power and sample size calculations for interaction

In planning a study in which interaction analyses may be of interest, it can be important to consider issues of power and sample size. Sample size and power calculations have been considered for multiplicative interaction using logistic regression (Hwang et al., 1994; Foppa and Spiegelman, 1997; Garcia-Closas and Lubin, 1999; Gauderman, 2002a; Demidenko, 2008), for case-only estimators of interaction (Yang et al., 1997; VanderWeele, 2011c), for additive interaction (VanderWeele, 2012a), and for multiplicative interaction using matched case–control data (Gauderman, 2002b). Software is available to implement a number of these power and sample size calculations. Windows-based, QUANTO, developed by Gauderman is available at http://hydra.usc.edu/gxe and will implement sample size calculations for likelihood ratio-based tests of interaction using various study designs. An Excel spreadsheet that can be used for sample size and power calculations for additive interaction, as well as for multiplicative interaction (on the risk ratio or odds scale or using a case-only estimator), for cohort or case–control data is given in VanderWeele (2012a). Appendix 2 of that paper provides a guide to the use of these spreadsheets.

A few patterns also merit comment and can be useful to consider when planning studies for interaction. First, in general, larger sample sizes are needed to be able to detect significant interaction than to simply detect significant overall effects. Second, when the independence assumption holds, the case-only estimator of multiplicative interaction is more powerful than the estimator from logistic regression (Yang et al., 1997). Third, for the classical interaction pattern of positive main effects for both exposures and positive interaction, the test for additive interaction is in general more powerful than the test for multiplicative interaction (Greenland, 1983; VanderWeele, 2012a).

3 Conclusions

In this tutorial, we have provided an introduction to the measures of, estimation procedures for, and interpretation relevant to interaction analyses. We have considered both additive and multiplicative measures and discussed the relative merits of each, as well as their relation to statistical models, along with case-only estimators to estimate multiplicative interaction. We have discussed confounding control and the interpretation of interaction analyses. We have also discussed the stronger conditions which are needed for a mechanistic interpretation of interactions. We have commented on extensions to continuous outcomes, on qualitative interaction, and on the informative presentation of interaction analyses and have given a brief summary of resources available for sample size and power calculations for interaction analyses.

There are a number of issues that we have not been able to touch upon in this tutorial. We have focused here on binary outcomes. Similar issues concerning additive versus multiplicative interaction are also relevant for time-to-event outcomes: Li and Chambless (2007) discussed additive interaction for the proportional hazard models; VanderWeele (2011b) discussed mechanistic interpretation of such additive interactions in time-to-event models; Rod et al. (2012) discussed interaction analysis in additive hazard models. Some of the recent research on interaction concern methods to robustly estimate interaction even if models for the main effects are misspecified (Vansteelandt et al., 2008, 2012; Tchetgen Tchetgen, 2010; Tchetgen Tchetgen and Robins, 2010). Another group of papers has examined methods to try to better exploit the conditional independence assumption of the case-only estimator when data are also available on controls (Chatterjee and Carroll, 2005; Mukherjee et al., 2007; Han et al., 2012) or methods that attempt to exploit the conditional independence assumption while still being at least partially protected against possible violations of this assumption (Mukherjee and Chatterjee, 2008; Dai et al., 2012). Methods are also available to jointly test a main effect and an interaction (Chatterjee et al., 2006; Kraft et al., 2007; Maity et al., 2009) so as to attempt to leverage potential interaction to be able to more powerfully detect genetic associations. Other work has examined methods to estimate interaction in family-based genetic studies design (Umbach and Weinberg, 2000; Lake and Laird, 2004; Hoffmann et al., 2009; Weinberg et al., 2011). Recently there has also been considerable interest in the challenges of assessing interaction in genome-wide-association studies when multiple comparison problems are present (Kraft, 2004; Gayan et al., 2008; Khoury and Wacholder, 2009; Murcray et al., 2009; Pierce and Ahsan, 2010; Thomas, 2010). In some settings exposures may vary over time and new methods have been developing to assess effect modification by time-varying covariates and/or exposures (Petersen et al., 2007; Robins et al., 2007; VanderWeele et al., 2010; Almirall et al., 2010). Further literature has noted that in many settings at least when the two exposures are independent in distribution, interaction may be robust to measurement error (Garcia-Closas et al., 1998; Zhang et al., 2008; Cheng and Lin, 2009; Lindström et al., 2009; Tchetgen Tchetgen and Kraft, 2011; VanderWeele, 2012b) even when such sources of bias render estimates of main effect invalid. Some of these topics are described in textbook form elsewhere (VanderWeele, 2014b). We have not been able to describe all of these methods and developments in this paper but we hope that the interested reader will consult the relevant literature and we hope also that this tutorial has provided a useful introduction to how to carry out and interpret analyses of interaction.

Appendix 1: SAS code for additive interaction estimates and confidence intervals

SAS code for additive interaction for binary exposures

Suppose we have a dataset named "mydata" with outcome variable "d", exposure variables "g" and "e", and three covariates "c1", "c2", and "c3". To calculate the relative excess risk due to interaction we can run a standard logistic regression in SAS using proc logistic where we add "outest = myoutput covout" to the procedure statement and then we also run the code that follows. The output will include the estimate of

RERI, its standard error, and a 95% confidence interval. Note that the first three independent variables in the model statement must be the two exposures, and their interaction or the code will not work (the other covariates can be entered in any order). Note also that if the class statement is used for proc logistic for categorical confounders, the exposures must NOT be included in the class statement or it will reverse the coding of the exposures and get the wrong results.

```
proc logistic descending data=mydata outest=myoutput covout;
  model d=g e g*e c1 c2 c3;
run;

data rerioutput;
  set myoutput;
  array mm {*} _numeric_;
  b0=lag4(mm[1]);
  b1=lag4(mm[2]);
  b2=lag4(mm[3]);
  b3=lag4(mm[4]);
  v11=lag2(mm[2]);
  v12=lag(mm[2]);
  v13=mm[2];
  v22=lag(mm[3]);
  v23=mm[3];
  v33=mm[4];
  k1=exp(b1+b2+b3)-exp(b1);
  k2=exp(b1+b2+b3)-exp(b2);
  k3=exp(b1+b2+b3);
  vreri=v11*k1*k1+v22*k2*k2+ v33*k3*k3+2*v12*k1*k2+2*v13*k1*k3
+ 2*v23*k2*k3;
  reri=exp(b1+b2+b3)-exp(b1)-exp(b2)+1;
  se_reri=sqrt(vreri);
  ci95_l=reri-1.96*se_reri;
  ci95_u=reri+1.96*se_reri;
  keep reri se_reri ci95_l ci95_u;
  if _n_=5;
run;

proc print data=rerioutput;
  var reri se_reri ci95_l ci95_u;
run;
```

SAS code for additive interaction for ordinal and continuous exposures

We can adapt this code also to calculate *RERI* for exposures which are ordinal or continuous. Suppose we wish to calculate the relative excess risk due to interaction comparing two different levels of the first exposure "g", say level 0 to level 2, and two different levels of our second exposure "e", say level 5 to level 25. We could then use the code below. Mathematical justification is given in the online supplement to this tutorial. In the code below the user must input the two levels being compared for both exposures at the beginning of the data step, e.g. "g1 = 2; g0 = 0; e1 = 25; e0 = 5;" or whatever values are of interest in

comparing. Note that if the user fixes "g1 = 1; g0 = 0; e1 = 1; e0 = 0;" then this will give the same output as the previous code above for binary exposures. Note that the first three independent variables in the model statement must be the two exposures and their interaction or the code will not work (the other covariates can be entered in any order). Note also that if the class statement is used for proc logistic for categorical confounders, the exposures must NOT be included in the class statement or it will reverse the coding of the exposures and get the wrong results.

```
proc logistic descending data=mydata outest=myoutput covout;
   model d=g e g*e c1 c2 c3;
run;

data rerioutput;
   set myoutput;
   g1=2;
   g0=0;
   e1=25;
   e0=5;
   array mm {*} _numeric_;
   b0=lag4(mm[1]);
   b1=lag4(mm[2]);
   b2=lag4(mm[3]);
   b3=lag4(mm[4]);
   v11=lag2(mm[2]);
   v12=lag(mm[2]);
   v13=mm[2];
   v22=lag(mm[3]);
   v23=mm[3];
   v33=mm[4];
   k1=(g1-g0)*exp((g1-g0)*b1+(e1-e0)*b2+(g1*e1-g0*e0)*b3)
-(g1-g0)*exp((g1-g0)*b1+(g1-g0)*e0*b3);
   k2=(e1-e0)*exp((g1-g0)*b1+(e1-e0)*b2+(g1*e1-g0*e0)*b3)
-(e1-e0)*exp((e1-e0)*b2+(e1-e0)*g0*b3);
   k3=(g1*e1-g0*e0)*exp((g1-g0)*b1+(e1-e0)*b2+(g1*e1-g0*e0)*b3)
-(g1-g0)*e0*exp((g1-g0)*b1+(g1-g0)*e0*b3)
-(e1-e0)*g0*exp((e1-e0)*b2+(e1-e0)*g0*b3);
   vreri=v11*k1*k1+v22*k2*k2+v33*k3*k3+2*v12*k1*k2+2*v13*k1*k3
+ 2*v23*k2*k3;
   reri=exp((g1-g0)*b1+(e1-e0)*b2+(g1*e1-g0*e0)*b3)
-exp((g1-g0)*b1+(g1-g0)*e0*b3)
-exp((e1-e0)*b2+(e1-e0)*g0*b3)+1;
   se_reri=sqrt(vreri);
   ci95_l=reri-1.96*se_reri;
   ci95_u=reri+1.96*se_reri;
   keep reri se_reri ci95_l ci95_u;
   if _n_=5;
run;

proc print data=rerioutput;
  var reri se_reri ci95_l ci95_u;
run;
```

SAS code for additive interaction for categorical exposures

For categorical exposures, to obtain estimates and confidence intervals for additive interaction one can restrict attention to two specific levels of each of the two variables and calculate measures of additive interaction using the code for binary exposures above. It is possible to proceed in this manner for each possible comparison of two levels of each of the two exposures. For example, if there were two categorical variables, A and B, and A had three levels (A1, A2, and A3) and B had four levels (B1, B2, B3, and B4), then one could assess additive interaction comparing A = A1 and A = A2 and B = B1 and B = B4 by ignoring the observations with A = A3 and also ignoring those with B = B2 or B = B3 and then using the code for binary exposures above. Suppose the name of the dataset with the categorical variables was mycatdata. We could then use the following SAS code:

```
data mydata;
   set mycatdata;
   if A = 'A1' then g = 0;
   if A = 'A2' then g = 1;
   if B = 'B1' then e = 0;
   if B = 'B4' then e = 1;
   if A = 'A1' or A = 'A2';
   if B = 'B1' or B = 'B4';
run;
```

The code deletes the observations with A = A3 and those with B = B2 or B = B3 and creates a new dataset only with values of A which are A1 or A2 and with values of B which are B1 or B4. The code for additive interaction for binary exposures can then be used directly. We could similarly proceed with any other comparison. We could compare (A1,A2) and (B1,B2); or (A1,A2) and (B1,B3); or (A1,A3) and (B1,B2); and so on.

Appendix 2: Stata code for additive interaction estimates and confidence intervals

Stata code for additive interaction for binary exposures

Suppose we have a dataset with outcome variable "d", exposure variables "g" and "e", and three covariates "c1", "c2", and "c3". To calculate the relative excess risk due to interaction we can: create an interaction variable "Ige", then run a standard logistic regression in Stata using the logit command, and then use Stata "nlcom" command in the code that follows. The output will include the estimate of *RERI*, its standard error, and a 95% confidence interval.

```
generate Ige = g*e
logit d g e Ige c1 c2 c3
nlcom exp(_b[g] + _b[e] + _b[Ige]) - exp(_b[g]) - exp(_b[e]) + 1
```

Stata code for additive interaction for ordinal and continuous exposures

We can also calculate *RERI* using Stata for exposures which are ordinal or continuous. Suppose we wish to calculate the relative excess risk due to interaction comparing two different levels of the first exposure "g",

say level 0 to level 2, and two different levels of our second exposure "e", say level 5 to level 25. We could then use the code below. In this code the user must specify, in the first four lines of code, the levels of both exposures that are being compared (in the code below the two levels for "g" are 2 and 0 and the two levels for "e" are "25" and "5" but these can be changed). If the user fixes g1 = 1; g0 = 0; e1 = 1; and e0 = 0, then the code will give the same output as the previous code above for binary exposures. The next two lines of code generate an interaction variable between "g" and "e" and fit the logistic regression model allowing for interaction. The final line of code uses the "nlcom" command in Stata to obtain *RERI*. The output will include the estimate of *RERI*, its standard error, and a 95% confidence interval.

```
generate g1=2
generate g0=0

generate e1=25
generate e0=5

generate Ige=g*e
logit d g e Ige c1 c2 c3

nlcom exp((g1-g0)*_b[g]+(e1-e0)*_b[e]+(g1*e1-g0*e0)*_b[Ige])
   -exp((g1-g0)*_b[g]+(g1-g0)*e0*_b[Ige])-exp((e1-e0)*_b[e]+
   (e1-e0)*g0*_b[Ige])+1
```

Stata code for additive interaction for categorical exposures

For categorical exposures, to obtain estimates and confidence intervals for additive interaction one can restrict attention to two specific levels of each of the two variables and calculate measures of additive interaction using the code for binary exposures above. It is possible to proceed in this manner for each possible comparison of two levels of each of the two exposures. For example, if there were two categorical variables, A and B, and A had three levels (A1, A2, A3) and B had four levels (B1, B2, B3, B4), then one could assess additive interaction comparing A = A1 and A = A2, and B = B1 and B = B4, by ignoring the observations with A = A3 and also ignoring those with B = B2 or B = B3 and then using the code for binary exposures above. We could create the restricted dataset using the following Stata code:

```
generate g=0 if A== 'A1';
replace g=1 if A== 'A2';
generate e=0 if B== 'A1';
replace e=1 if B== 'B4';
```

The code for additive interaction for binary exposures can then be used directly. The code creates variables g and e only for those the observations with values of A which are A1 or A2 and with values of B which are B1 or B4. When the code for additive interaction for binary exposures is used it will only analyze the observations with values of A which are A1 or A2 and with values of B which are B1 or B4 since those with values of A which are A3 or with values of B which are B2 or B3 will have their values of g and of e missing.

We could similarly proceed with any other comparison. We could compare (A1,A2) and (B1, B2); or (A1, A2) and (B1, B3); or (A1, A3) and (B1, B2); and so on.

References

Ai, C. and Norton, E. C. (2003). Interaction terms in logit and probit models. Economics Letters, 80:123–129.

Albert, P. S., Ratnasinghe, D., Tangrea, J., and Wacholder, S. (2001). Limitations of the case-only design for identifying gene–environment interactions. American Journal of Epidemiology, 154:687–693.

Almirall, D., Ten Have, T., and Murphy, S. A. (2010). Structural nested mean models for assessing time-varying effect moderation. Biometrics, 66:131–139.

Andersson, T., Alfredsson, L., Kallberg, H., Zdravkovic, S., and Ahlbom, A. (2005). Calculating measures of biological interaction. European Journal of Epidemiology, 20:575–579.

Assmann, S. F., Hosmer, D. W., Lemeshow, S., and Mundt, K. A. (1996). Confidence intervals for measures of interaction. Epidemiology, 7:286–290.

Bennett, W. P., Alavanja, M. C. R., Blomeke, B., Vähäkangas, K. H., Castrén, K., Welsh, J. A., Bowman, E. D., Khan, M. A., Flieder, D. B., and Harris, C. C. (1999). Environmental tobacco smoke, genetic susceptibility, and risk of lung cancer in never-smoking women. Journal of the National Cancer Institute, 91:2009–2014.

Bhavnani, D., Goldstick, J. E., Cevallos, W., Trueba, G., and Eisenberg, J. N. S. (2012). Synergistic effects between rotavirus and coinfecting pathogens on diarrheal disease: Evidence from a community-based study in northwestern Ecuador. American Journal of Epidemiology, 176:387–395.

Blot, W. J. and Day, N. E. (1979). Synergism and interaction: Are they equivalent? American Journal of Epidemiology, 110:99–100.

Bonetti, M. and Gelber, R. D. (2000). A graphical method to assess treatment-covariate interactions using the cox model on subsets of the data. Statistics in Medicine, 19:2595–2609.

Bonetti, M. and Gelber, R. D. (2005). Patterns of treatment effects in subsets of patients in clinical trials. Biostatistics, 5:465–481.

Botto, L. D. and Khoury, M. J. (2001). Facing the challenge of gene–environment interaction: the two-by-four table and beyond. American Journal of Epidemiology, 153:1016–1020.

Cai, T., Tian, L., Wong, P. H., and Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics, 12:270–282.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene–environment independence in case–control studies. Biometrika, 92:399–418.

Chatterjee, N., Kalaylioglu, Z., Moleshi, R., Peters, U., and Wacholder, S. (2006). Powerful multilocus tests of genetic association in the presence of gene–gene and gene–environment interactions. American Journal of Human Genetics, 79:1002–1016.

Cheng, K. F. and Lin, W. J. (2009). The effects of misclassification in studies of gene–environment interactions. Human Heredity, 67:77–87.

Chu, H., Nie, L., and Cole, S. R. (2011). Estimating the relative excess risk due to interaction: A Bayesian approach. Epidemiology, 22:242–248.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics, 11, 2463–2468.

Cordell, H. J. (2009). Detecting gene–gene interaction that underlie human diseases. Nature Reviews Genetics, 10:392–404.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, L. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. Journal of the National Cancer Institute, 22:173–203.

Dai, J., Logsdon, B., Huang, Y., et al. (2012). Simultaneous testing for marginal genetic association and gene–environment interaction in genome-wide association studies. American Journal of Epidemiology, 176:164–173.

de González, A. B, and Cox, D. R. (2007). Interpretation of interaction: A review. Annals of Applied Statistics, 1:371–385.

Deeks, J. J. and Altman, D. G. (2003). Effect measures for met-analysis of trials with binary outcomes. In: Systematic Reviews in Health Care: Meta-Analysis in Context, M. Egger, G. Davey Smith, and D. G. Altman (Eds.), 313–335. London: BMJ Publishing Group.

Demidenko, E. (2008). Sample size and optimal design for logistic regression with binary interaction. Statistics in Medicine, 27:36–46.

Engels, E. A., Schmid, C. H., Terrin, N., et al. (2000). Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. Statistics in Medicine, 19:1707–1728.

Figueiredo, J. C., Knight, J. A., Briollais, L., Andrulis, I. L., and Ozcelik, H. (2004). Polymorphisms XRCC1-R399Q and XRCC3-T241M and the risk of breast cancer at the Ontario Site of the Breast Cancer Family Registry. Cancer Epidemiology, Biomarkers and Prevention, 13:583–591.

Foppa, I. and Spiegelman, D. (1997). Power and sample size calculations for case–control studies of gene–environment interactions with a polytomous exposure variable. American Journal of Epidemiology, 146:596–604.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. Biometrics, 41:361–372.

Garcia-Closas, M. and Lubin, J. H. (1999). Power and sample size calculations in case–control studies of gene–environment interactions: Comments on different approaches. American Journal of Epidemiology, 149:689–692.

Garcia-Closas, M., Thompson, W. D., and Robins, J. M. (1998). Differential misclassification and the assessment of gene–environment interactions. American Journal of Epidemiology, 147:426–433.

Gauderman, W. J. (2002a). Sample size requirements for association studies of gene–gene interaction. American Journal of Epidemiology, 155:478–484.

Gauderman, W. J. (2002b). Sample size requirements for matched case–control studies of gene–environment interaction. Statistics in Medicine, 21:35–50.

Gayan, J., et al. (2008). A method for detecting epistasis in genome-wide studies using case–control multi-locus association analysis. BMC Genomics, 9:360.

Greenland, S. (1983). Tests for interaction in epidemiologic studies: A review and study of power. Statistics in Medicine, 2:243–251.

Greenland, S. (2009). Interactions in epidemiology: relevance, identification and estimation. Epidemiology, 20:14–17.

Greenland, S., Lash, T. L., and Rothman, K. J. (2008). "Concepts of interaction," chapter 5. In: Modern Epidemiology, K. J. Rothman, S. Greenland, and T. L. Lash (Eds.). 3rd Edition. Philadelphia, PA: Lippincott Williams and Wilkins.

Han, S. S., Rosenberg, P. S., Garcia-Closas, M., Figueroa, J. D., Silverman, D., Chanock, S. J., Rothman, N., and Chatterjee, N. (2012). Likelihood ratio test for detecting gene (G)–environment (E) interactions under an additive risk model exploiting G-E independence for case–control data. American Journal of Epidemiology, 176:1060–1067.

Hill, A. B. (1965). The environment and disease: Association or causation? Proceedings of the Royal Society of Medicine, 58:295–300.

Hilt, B., Langård, S., Lund-Larsen, P. G., and Lien, J. T. (1986). Previous asbestos exposure and smoking habits in the county of Telemark, Norway – A cross-sectional population study. Scandinavian Journal of Work, Environment and Health, 12:561–566.

Hoffmann, T. J., Lange, C., Vansteelandt, S., and Laird, N. M. (2009). Gene–environment interaction tests for dichotomous traits in trios and sibships. Genetic Epidemiology, 33:691–699.

Hosmer, D. W. and Lemeshow, S. (1992). Confidence interval estimation of interaction. Epidemiology, 3:452–456.

Hwang, S.-J., Beaty, T. H., Liang, K.-Y., Coresh, J., and Khoury, M. J. (1994). Minimum sample size estimation to detect gene–environment interaction in case–control designs. American Journal of Epidemiology, 140:1029–1037.

Khoury, M. J. and Wacholder, S. (2009). From Genome-wide association studies to gene–environment-wide interaction studies – Challenges and opportunities. American Journal of Epidemiology, 169:227–230.

Knol, M. J., Egger, M., Scott, P., Geerlings, M. I., and Vandenbroucke, J. P. (2009). When one depends on the other: Reporting of interaction in case–control and cohort studies. Epidemiology, 2009(20):161–166.

Knol, M. J., Vandenbroucke, J. P., Scott, P., and Egger, M. (2008). What do case–control studies estimate? Survey of methods and assumptions in published case–control research. American Journal of Epidemiology, 168:1073–1081.

Knol, M. J. and VanderWeele, T. J. (2012). Guidelines for presenting analyses of effect modification and interaction. International Journal of Epidemiology, 41:514–520.

Knol, M. J., VanderWeele, T. J., Groenwold, R. H. H., Klungel, O. H., Rovers, M. M., and Grobbee, D. E. (2011). Estimating measures of interaction on an additive scale for preventive exposures. European Journal of Epidemiology, 26:433–438.

Knol, M. J., le Cessie, S., Algra, A., Vandenbroucke, J. P., and Groenwold, R. H. H. (2012). Overestimation of risk ratios by odds ratios in trials and cohort studies: Alternatives to logistic regression. Canadian Medical Association Journal, 184:895–899.

Knol, M. J., van der Tweel, I., Grobbee, D. E., Numans, M. E., and Geerlings, M. I. (2007). Estimating interaction on an additive scale between continuous determinants in a logistic regression model. International Journal of Epidemiology, 36:1111–1118.

Kraft, P. (2004). Multiple comparisons in studies of gene x gene and gene x environment interaction. American Journal of Human Genetics, 74:582–585.

Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J., and Gauderman, W. J. (2007). Exploiting gene–environment interaction to detect disease susceptibility loci. Human Heredity, 63:111–119.

Kuss, O., Schmidt-Pokrzywniak, A., and Stang, A. (2010). Confidence intervals for the interaction contrast ratio. Epidemiology, 21:273–274.

Kuyvenhoven, J. P., Veenendaal, R. A., and Vandenbroucke, J. P. (1999). Peptic ulcer bleeding: Interaction between non-steroidal anti-inflammatory drugs, Helicobacter pylori infection, and the ABO blood group system. Scandinavian Journal of Gastroenterol, 34:1082–1086.

Lake, S. and Laird, N. (2004). Tests of gene–environment interaction for case-parent triads with general environmental exposures. Annals of Human Genetics, 68:55–64.

Lawlor, D. A. (2011). Biological interaction: Time to drop the term? Epidemiology, 22:148–150.

Li, Y., et al. (2010). Genetic variants and risk of lung cancer in never smokers: A genome-wide association study. Lancet Oncology, 11:321–330.

Li, R. and Chambless, L. (2007). Test for additive interaction in proportional hazards models. Annals of Epidemiology, 17:227–236.
Li, J. and Chan, I. S. (2006). Detecting qualitative interactions in clinical trials: An extension of range test. Journal of Biopharmaceutical Statistics, 16:831–841.
Lindström, S., Yen, Y.-C., Spiegelman, D., and Kraft, P. (2009). The impact of gene–environment dependence and misclassification in genetic association studies incorporating gene–environment interactions. Human Heredity, 68:171–181.
Lundberg, M., Fredlund, P., Hallqvist, J., and Diderichsen, F. (1996). A SAS program calculating three measures of interaction with confidence intervals. Epidemiology, 7:655–656.
Maity, A., Carroll, R. J., Mammen, E., and Chatterjee, N. (2009). Testing in semiparametric models with interaction, with applications to gene–environment interactions. Journal of the Royal Statistical Society, Series B, 71:75–96.
Miller, D. P., Liu, G., De Vivo, I., et al. (2002). Combinations of the variant genotypes of GSTP1, GSTM1, and p53 are associated with an increased lung cancer risk. Cancer research, 62:2819–2823.
Mukherjee, B. and Chatterjee, N. (2008). Exploiting gene–environment independence for analysis of case–control studies: An empirical-Bayes type shrinkage estimator to trade off between bias and efficiency. Biometrics, 64:685–694.
Mukherjee, B., Zhang, L., Ghosh, M., and Sinha, S. (2007). Semiparametric Bayesian analysis of case–control data under conditional gene–environment independence. Biometrics, 63:834–844.
Murcray, C. E., Lewinger J. P., and Gauderman, W. J. (2009). Gene–environment interaction in genome-wide association studies. American Journal of Epidemiology, 169:219–226.
Nie, L., Chu, H., Li, F., and Cole, S. R. (2010). Relative excess risk due to interaction: resampling-based confidence intervals. Epidemiology, 21:552–556.
Norton, E. C., Wang, H., and Ai, C. (2004). Computing interaction effects and standard errors in logit and probit models. Stata Journal, 4:154–167.
Pan, G. and Wolfe, D. A. (1997). Test for qualitative interaction of clinical significance. Statistics in Medicine, 16:1645–1652.
Petersen, M. L., Deeks, S. G., Martin, J. N., and van der Laan, M. J. (2007). History-adjusted marginal structural models for estimating time-varying effect modification. American Journal of Epidemiology, 166:985–993.
Peto, R. (1982). Statistical aspects of cancer trials. In: Treatment of Cancer, K. E. Halnan (Ed.), 867–871. London: Chapman and Hall.
Phillips, P. C. (2008). Epistasis – The essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetic, 9:855–867.
Piantadosi, S. and Gail, M. H. (1993). A comparison of the power of two tests for qualitative interactions. Statistics in Medicine, 12:1239–1248.
Piegorsch, W. W., Weinberg, C. R., and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case–control studies. Statistics in Medicine, 13:153–162.
Pierce, B. L. and Ahsan, H. (2010). Case-only genome-wide interaction study of disease risk, prognosis and treatment. Genetic Epidemiology, 34:7–15.
Poole, C. (2010). On the origin of risk relativism. Epidemiology, 21:3–9.
Richardson, D. B. and Kaufman, J. S. (2009). Estimation of the relative excess risk due to interaction and associated confidence bounds. American Journal of Epidemiology, 169:756–760.
Robins, J. M.Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology, 11:550–560.
Robins, J. M., Hernán, M. A., and Rotnitzky, A. (2007). Effect modification by time-varying covariates. American Journal of Epidemiology, 166:994–1002.
Rod, N. H., Lange, T., Andersen, I., Marott, J. L., Diderichsen, F. (2012). Additive interaction in survival analysis: use of the additive hazards model. Epidemiology. 23:733–737.
Rothman, K. J. (1976). Causes. American Journal of Epidemiology, 104:587–592.
Rothman, K. J. (1986). Modern Epidemiology. 1st Edition. Boston, MA: Little, Brown and Company.
Rothman, K. J., Greenland, S., and Walker, A. M. (1980). Concepts of interaction. American Journal of Epidemiology, 112:467–470.
Rothman, K. J., and Greenland, S. editors. (1998). Modern epidemiology. 2nd Edition. Philadelphia: Lippincott.
Saracci, R. (1980). Interaction and synergism. American Journal of Epidemiology, 112:465–466.
Siemiatycki, J. and Thomas, D. C. (1981). Biological models and statistical interactions: An example from multistage carcinogenesis. International Journal of Epidemiology, 10:383–387.
Silvapulle, M. J. (2001). Tests against qualitative interaction: Exact critical values and robust tests. Biometrics, 57:1157–1165.
Skrondal, A. (2003). Interaction as departure from additivity in case–control studies: A cautionary note. American Journal of Epidemiology, 158(3):251–258.
Song, X. and Pepe, M. S. (2004). Evaluating markers for selecting a patient's treatment. Biometrics, 60:874–883.
Sterne, J. A. and Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. Journal of Clinical Epidemiology, 54:1046–1055.
Szklo, M. and Nieto, F. J. (2007). Epidemiology: Beyond the Basics. 2nd Edition. Boston, MA: Jones and Bartlee Publishers.

Tchetgen Tchetgen, E. J. (2010). On the interpretation, robustness, and power of varieties of case-only tests of gene–environment interaction. American Journal of Epidemiology, 172:1335–1338.
Tchetgen Tchetgen, E. J. and Kraft, P. (2011). On the robustness of tests of genetic associations incorporating gene–environment interaction when the environmental exposure is misspecified. Epidemiology, 22:257–261.
Tchetgen Tchetgen, E. J. and Robins, J. M. (2010). The semi-parametric case-only estimator. Biometrics, 66:1138–1144.
Tchetgen Tchetgen, E. J. and VanderWeele, T. J. (2012). Robustness of measures of interaction to unmeasured confounding. Harvard University, Technical Report.
Thomas, D. (2010). Gene–environment-wide association studies: Emerging approaches. Nature Reviews Genetics, 11:259–272.
Thompson, W. D. (1991). Effect modification and the limits of biologic inference from epidemiologic data. Journal of Clinical Epidemiology, 44:221–232.
Umbach, D. and Weinberg, C. (2000). The use of case-parent triads to study joint effects of genotype and exposure. American Journal of Human Genetics, 66:251–261.
Vandenbroucke, J. P., Koster, T., Briët, E., Reitsma, P. H., Bertina, R. M., and Rosendaal, F. R. (1994). Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. Lancet, 344:1453–1457.
Vandenbroucke, J. P., von Elm, E., Altman, D. G., et al. (2007). Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. Epidemiology, 18:805–835.
VanderWeele, T. J. (2009a). On the distinction between interaction and effect modification. Epidemiology, 20:863–871.
VanderWeele, T. J. (2009b). Sufficient cause interactions and statistical interactions. Epidemiology, 20:6–13.
VanderWeele, T. J. (2010a). Empirical tests for compositional epistasis. Nature Reviews Genetics, 11:166.
VanderWeele, T. J. (2010b). Epistatic interactions. Statistical Applications in Genetics and Molecular Biology, 9(Article 1):1–22.
VanderWeele, T. J. (2010c). Response to "On the definition of effect modification," by E. Shahar and D.J. Shahar. Epidemiology, 21:587–588.
VanderWeele, T. J. (2010d). Sufficient cause interactions for categorical and ordinal exposures with three levels. Biometrika, 97:647–659.
VanderWeele, T. J. (2011a). A word and that to which it once referred: assessing "biologic" interaction. Epidemiology, 22:612–613.
VanderWeele, T. J. (2011b). Causal interactions in the proportional hazards model. Epidemiology, 22:713–717.
VanderWeele, T. J. (2011c). Sample size and power calculations for case-only interaction studies: Formulas for common test statistics. Epidemiology, 22:873–874.
VanderWeele, T. J. (2012a). Sample size and power calculations for additive interactions. Epidemiologic Methods, 1:159–188.
VanderWeele, T. J. (2012b). Interaction tests under exposure misclassification. Biometrika, 99:502–508.
VanderWeele, T. J. (2013). Reconsidering the denominator of the attributable proportion for additive interaction. European Journal of Epidemiology, 28:779–784.
VanderWeele, T. J. (2014a). A unification of mediation and interaction: A four-way decomposition. Epidemiology, in press.
VanderWeele, T. J. (2014b). Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford University Press, in press.
VanderWeele, T. J. and Knol, M. J. (2011a). The interpretation of subgroup analyses in randomized trials: Heterogeneity versus secondary interventions. Annals of Internal Medicine, 154:680–683.
VanderWeele, T. J. and Knol, M. J. (2011b). Remarks on antagonism. American Journal of Epidemiology, 173:1140–1147.
VanderWeele, T. J., Mukherjee, B., and Chen, J. (2012). Sensitivity analysis for interactions under unmeasured confounding. *Statistics in Medicine*, 31:2552–2564.
VanderWeele, T. J. and Richardson, T. S. (2012). General theory for interactions in sufficient cause models with dichotomous exposures. Annals of Statistics, 40:2128–2161.
VanderWeele, T. J. and Robins, J. M. (2007a). The identification of synergism in the SCC framework. Epidemiology, 18:329–339.
VanderWeele, T. J. and Robins, J. M. (2007b). Four types of effect modification – A classification based on directed acyclic graphs. Epidemiology, 18:561–568.
VanderWeele, T. J. and Robins, J. M. (2008). Empirical and counterfactual conditions for sufficient cause interactions. Biometrika, 95:49–61.
VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2014). Attributing effects to interactions. Epidemiology, in press.
VanderWeele, T. J. and Vansteelandt, S. (2011). A weighting approach to causal effects and additive interaction in case–control studies: Marginal structural linear odds models. American Journal of Epidemiology, 174:1197–1203.
VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2010). Marginal structural models for sufficient cause interactions. American Journal of Epidemiology, 171:506–514.
Vansteelandt, S., VanderWeele, T. J., and Robins, J. M. (2012). Semiparametric inference for sufficient cause interactions. Journal of the Royal Statistical Society, Series B, 74:223–244.
Vansteelandt, S., VanderWeele, T. J., Tchetgen, E. J., and Robins, J. M. (2008). Multiply robust inference for statistical interactions. Journal of the American Statistical Association, 103:1693–1704.
Walter S. D., and Holford, T. R. (1978). Additive, multiplicative, and other models for disease risks. American Journal of Epidemiology, 108:341–346.

Weinberg, C. R., Shi, M., and Umbach, D. M. (2011). A sibling-augmented case-only approach for assessing multiplicative gene–environment interactions. American Journal of Epidemiology, 174:1183–1189.

Yang, Q., Khoury, M. J., and Flanders, W. D. (1997). Sample size requirements in case-only designs to detect gene–environment interaction. American Journal of Epidemiology, 146:713–719.

Yang, Q., Khoury, M. J., Sun, F., and Flanders, W. D. (1999). Case-only design to measure gene–gene interaction. Epidemiology, 10:167–170.

Yelland, L. N., Salter, A. B., and Ryan, P. (2011). Relative risk estimation in randomized controlled trials: a comparison of methods for independent observations. International Journal of Biostatistics, 7(1):1–31.

Zhang, L., Mukherjee, B., Ghosh, M., Gruber, S., and Moreno, V. (2008). Accounting for error due to misclassification of exposures in case–control studies of gene–environment interaction. Statistics in Medicine, 27:2756–2783.

Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L. J. (2013). Effectively selecting a target population for a future comparative study. Journal of the American Statistical Association, 108:527–539.

Zou, G. Y. (2008). On the estimation of additive interaction by use of the four-by-two table and beyond. American Journal of Epidemiology, 168:212–224.