Choosing Statistical Models to Assess Biological Interaction as a Departure from Additivity of Effects

David M. Thompson, Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104

Yan Daniel Zhao, Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104

Key Words: additivity and multiplicativity of effects; biological interaction; statistical interaction; generalized linear models; interaction contrast; Relative Excess Risk Due to Interaction (RERI)

Abstract

Vanderweele and Knol define biological interaction as an instance wherein "two exposures physically interact to bring about the outcome." A hallmark of biological interaction is that the total effect, produced when factors act together, differs from the sum of effects when the factors operate independently. Epidemiologists construct statistical models to assess biological interaction. A consensus exists that biological interaction should be assessed as a departure from additivity of effects.

This paper compares three statistical models' assessment of biological interaction in a data example that appears in several epidemiology textbooks. A linear binomial model quantifies a departure from additivity in the data example in terms of differences in probabilities. It generates directly interpretable estimates and 95% confidence intervals for parameters including the interaction contrast (IC). Log binomial and logistic regression models detect no departure from multiplicativity in the data example. However, their results permit calculation of the "Relative Excess Risk Due to Interaction" (RERI), a measure of departure from additivity on a relative risk scale.

The linear binomial model directly produces interpretable assessments of departures from additivity of effects and deserves wider use in research and in the teaching of epidemiology. Strategies exist to address the model's limitations.

1. Background

Hypotheses related to biological interaction are often of interest in studies of clinical or population health. Vanderweele and Knol (2014, p. 54) define biological interaction as an instance in which "two exposures physically interact to bring about the outcome." Rothman (2002, p. 171) states that "biologic interaction between two causes occurs whenever the effect of one is dependent on the presence of the other." Rothman's definition is closely allied with the concept of effect modification.

1.1 Biological interaction and statistical interaction

Investigators detect interaction and effect modification by constructing statistical models. Rothman (2002, p.169) points out that "in statistics, the term 'interaction' is used to refer to departure from the underlying form of a statistical model." Certain statistical models are suited for detecting departures from additivity of effects, and others are suited for detecting departures from multiplicativity of effects.

Researchers frequently hypothesize biological mechanisms that produce a non-additivity of effects when those effects are quantified as probabilities. Rothman links "biological independence" with an additivity of effects and connects "biological interaction" with a departure from an additivity of effects. "Why is it," Rothman asks, "that biological interaction should be evaluated as departures from additivity of effect" (Rothman, 2002, p. 178)? By 2007, the STROBE statement regarded the response to Rothman's rhetorical question to reflect a "consensus that the additive scale, which uses absolute risks, is more appropriate [than the multiplicative scale] for public health and clinical decision making" (Vandenbroucke, von Elm, et al., 2007, p.817). The authors of the STROBE statement remind investigators that "in many circumstances, the absolute risk associated with an exposure is of greater interest than the relative risk" and ask them to "consider translating estimates of relative risk into absolute risk for a meaningful time period" (p.825). Vanderweele and Knol (2014, p. 37) remark, more pointedly, that "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."

1.2 Additivity and multiplicativity of effects

Although this paper presents ideas that align with this consensus, it avoids using the term "additive interaction." Instead, we link the concept to statistical models that assess evidence of a *departure from additivity of effects*. Spiegelman and Hertzmark (2005) describe the "binomial model for the risk difference," which directly assesses departures from additivity of effects in terms of probabilities and differences in probabilities. This model is also called the "binomial regression model" (Cheung 2007; Bieler et al., 2010). Richardson et al. (2015) call it the "linear binomial model," the term which this paper uses.

In the linear binomial model, detection of statistical interaction constitutes direct evidence of a departure from additivity of effects. In contrast, the log binomial and logistic regression models assess additivity indirectly, when their estimates of relative risks or odds ratios are recombined to calculate statistics like the "Relative Excess Risk due to Interaction" (RERI).

Similarly, instead of using the term "multiplicative interaction," the paper links that concept to statistical models that assess evidence of *departures from multiplicativity of effects.* Log binomial models estimate these effects in terms of relative risks, also called risk ratios, prevalence ratios (Spiegelman and Hertzmark 2005; Richardson et al., 2015) or prevalence proportion ratios. Logistic regression models estimate effects in terms of odds and odds ratios. In the log binomial and logistic models, which employ log transformations of probabilities or of their corresponding odds, detection of statistical interaction constitutes direct evidence of a departure from multiplicativity among effects. Because the choice of statistical model affects the

interpretation of statistical interaction, Rothman (2002, p.170) prefers the term "effect measure modification" to "effect modification" or to "interaction."

1.3 Statistical models for binomial outcomes

The linear binomial, log binomial and logistic regression models are all examples of generalized linear models. Each treats the outcome as arising from a binomial distribution. Each features a linear predictor structured as a *sum of terms*. In this regard, all generalized linear models might be considered "additive." Accordingly, this paper does not refer to "additive or multiplicative models" but refers instead to statistical models that assess additivity or multiplicativity of effects.

While the three models link a binomial outcome to a linear predictor, they are distinguished by the link functions they employ in the generalized linear model framework. The linear binomial model uses the identity link, the log binomial model uses the log link, and the logistic regression model uses the logit link. Thus, the linear binomial model operates directly on probabilities, while the others apply log transformations of the probabilities or of their corresponding odds. Because each model estimates a different effect measure, they differ in their ability to detect statistical interaction in a collection of data.

Section 2 introduces a definition of additivity of effects that are quantified as probabilities and differences in probabilities. The additivity of two effects can be characterized as an equality of joint and independent effects and, equivalently, as a homogeneity of effects. Interaction, which reflects a departure from additivity, can be characterized as an inequality between joint and independent effects, and as heterogeneity among effects. We discuss two formal assessments of a departure from additivity: the interaction contrast, whose terms are probabilities, and the RERI, whose terms are relative risks.

Section 3 introduces a data example, a widely cited example of biological interaction (Hammond et al. 1979), which features a dichotomous outcome, lung cancer mortality, and two dichotomous risk factors, occupational exposure to asbestos and cigarette smoking. This section illustrates how a linear binomial model, which directly assesses additivity by generating estimates and testing hypotheses on probabilities and differences in probabilities, detects statistical interaction in these data. It then illustrates how, applied to the same data, the log binomial and logistic regression models, which assess multiplicativity of relative risks or of odds ratios, find no evidence of statistical interaction. The absence of statistical interaction does not point in this instance to the absence of biological interaction, but to a lack of departure from multiplicativity of effects.

Section 4 summarizes the three models' advantages and limitations for assessing additivity of effects. The RERI's wide use persists despite complications in its estimation, testing and interpretation. In comparison, the linear binomial model produces estimates of effects, including the interaction contrast, that are readily interpretable. The model can encounter problems with convergence, but strategies have been proposed to address those.

2. Assessing additivity of effects that are quantified as probabilities and differences in probabilities

Consider a comparison of the probability or "risk" of an outcome Y among individuals who are exposed or not exposed to one or both of two "risk factors," X and Z. Then p_{xz}, is a probability whose subscripts signify the probability or risk of the outcome Y at "levels" of X and Z. Table 1 illustrates these probabilities.

Table 1. Probabilities of an outcome (Y) at levels of two exposure or risk factors, X and Z

	Z=1 ("exposed to factor Z")	Z=0 ("not exposed to factor Z")
X=1 ("exposed to factor X")	p_{11}	p_{10}
X=0 ("not exposed to factor X")	p_{01}	p_{00}

2.1 Additivity defined as the equality of joint and independent effects

Rothman (2002, p.178) states that the following equation "establishes additivity as the definition of biological independence."

$$p_{11} - p_{00} = (p_{10} - p_{00}) + (p_{01} - p_{00}) \qquad \text{(Equation 1)}$$

Two exposures (X and Z) are biologically independent when the effect on Y of their joint and simultaneous effects ($p_{11} - p_{00}$) is equal to the sum of the separate and independent effects of X ($p_{10} - p_{00}$) and of Z ($p_{01} - p_{00}$).

A *departure* from additivity of effect, which Rothman considers evidence of biological interaction, is present if the joint and simultaneous effect of the two exposures differs from the sum of the effects of each exposure when considered separately. Szklo (2004, p. 186) similarly states that "interaction occurs when the observed joint effect of X and Z differs from that expected on the basis of their independent effects."

2.2 Additivity defined as a homogeneity of effects

Additivity can be defined equivalently as a homogeneity of effects. The terms of Equation 1 can be reordered to obtain

$$p_{11} - p_{01} = p_{10} - p_{00}\,, \qquad \text{(Equation 2)}$$

which shows that additivity implies a homogeneity of effects. Equation 2 states that the effect of X on Y is the same whether $Z = 1$ ($p_{11} - p_{01}$) or $Z = 0$ ($p_{10} - p_{00}$). Additivity (the absence of interaction) implies that measures of association between Y and X are homogenous (do not differ) at levels of Z.

Homogeneity of effects is reciprocal. We can rearrange the probabilities in equation 2 and express them as:

$$p_{11} - p_{10} = p_{01} - p_{00}. \qquad \text{(Equation 3)}$$

Equation 3 states that the effect of Z on Y is the same at all levels of X, that is, whether X=1 ($p_{11} - p_{10}$) or X=0 ($p_{01} - p_{00}$). When effects of X and Z are additive, the association between Y and X is homogenous at levels of Z, and the association between Y and Z is homogenous at levels of X.

When the effects of X and Z are not additive, interaction is present. The effect on Y of their joint and simultaneous effects is either larger, or smaller, than the sum of their separate and independent effects. Equivalently, the effects of either X or Z on Y are heterogenous at levels of the other variable.

Any of the three equations (1, 2 or 3) can be rearranged to arrive at any of the others. Assessing homogeneity of effects, or assessing the equality of joint and independent effects, are algebraically equivalent ways to describe additivity.

2.3 Assessing additivity of effects using probabilities: the interaction contrast (IC)

Another arrangement of the terms in equation (1) is

$$(1) \times p_{11} + (-1) \times p_{10} + (-1) \times p_{01} + (1) \times p_{00} = 0 \qquad \text{(Equation 4)}$$

This linear contrast formally tests the hypothesis that the effects on Y of X and Z are additive or, equivalently, that no interaction exists between X and Z. Rothman refers to equation 4 as the "interaction contrast" (IC). An appropriate statistical model can estimate the quantity on the left side of the equation, calculate its 95% confidence interval, and judge whether it differs from zero.

2.4 Assessing additivity of effects using ratios: the RERI

Reordering the terms in Equation 1 and dividing each by p_{00} yields:

$$p_{11}/p_{00} - p_{01}/p_{00} - p_{10}/p_{00} + 1 = 0.$$

Recognizing that these ratios of probabilities are relative risks (RR), we obtain:

$$RR_{11} - RR_{01} - RR_{10} + 1 = 0. \qquad \text{(Equation 5)}$$

Rothman (1986) named the quantity on the left side of equation 5 the "Relative Excess Risk due to Interaction" (RERI). Rothman and Greenland (1998) call it the "interaction contrast ratio" (ICR). Hosmer and Lemeshow (1992) define it as "the proportion of disease among those with both exposures that is attributable to their interaction." Equation 5's expression for the RERI can be restated: $RR_{11} - 1 = (RR_{10} - 1) + (RR_{01} - 1)$. This equation's left side reflects the relative risk of experiencing the outcome in those exposed to the joint or simultaneous effects of both X and Z. Its right side reflects the sum or additive effects of exposure to just one of the factors X and Z.

The algebraic equivalence between equations 1 and 5 validates the assessment of additivity of effects using estimates of relative risks. If equation 5 holds and the RERI is zero, we conclude that the effects of X and Z conform to additivity *on a relative risk scale*. Evidence of inequality suggests a departure from additivity of effects, that is, evidence of interaction or effect modification on the relative risk scale.

The STROBE statement advocated use of the RERI as a measure of departures from additivity of risk differences (Vandenbroucke, von Elm, et al., 2007, p.825). The RERI is commonly used in epidemiologic research to quantify departures from additivity that researchers regard as evidence of biological interaction.

Under certain conditions (when the outcome is relatively rare in all strata defined by levels of X and Z, or when incidence density sampling is used to include non-cases from the underlying population), then odds ratios will approximate relative risks, and analogous equations will apply to estimates of odds ratios: $OR_{11} - OR_{01} - OR_{10} + 1 = 0$, or equivalently, $OR_{11} - 1 = (OR_{10} - 1) + (OR_{01} - 1)$ (VanderWeele, 2013).

3. Data example: lung cancer mortality among workers with different exposures to asbestos and smoking

A study (Hammond et al.,1979) that compared the risk of mortality from lung cancer among 17,800 asbestos workers in the US, and also among 73,763 men who were not exposed to asbestos, is widely used in epidemiology textbooks and teaching. The study also recorded smoking status, and so participants displayed combinations of exposure to cigarette smoking and to asbestos.

Appendix A contains the SAS program that created a dataset that closely approximates the features of the published data. So that the dataset's risk probabilities (reported as lung cancer deaths per 100,000) reflect the published ones, a smoking prevalence of 0.28 was assumed for both the asbestos workers and for the comparison group of unexposed workers. The SAS program also produced Table 2.

Table 2. Lung cancer deaths (per 100,000 workers) among those with exposure to asbestos and/or cigarette smoking

	Asbestos Exposure	
Cigarette smoking	Asbestos Workers (n= 17800)	Comparison Group (n=73763)
Smokers	p_{11}=601.9	p_{10}= 121.1
Non-smokers	p_{01}= 54.6	p_{00}= 11.3

3.1 The data example illustrates a departure from additivity of effects

If the effects of asbestos exposure and cigarette smoking are additive, the expected effect of experiencing both exposures would equal the sum of the exposures' separate effects (Equation 1). Following the notation introduced in Table 1 to define $\hat{p}_{xz}$, where X denotes cigarette smoking (1 = smokers and 0 = nonsmokers) and Z denotes asbestos exposure (1=exposed and 0= not exposed), the estimated risk probabilities are:

$\hat{p}_{11} - \hat{p}_{00} = 601.9 - 11.3 = 590.6$ excess deaths per 100,000 people, attributable to joint effects of both exposures.

$\hat{p}_{10} - \hat{p}_{00} = 121.0 - 11.3 = 109.7$ excess deaths per 100,000 attributable to smoking by itself;

$\hat{p}_{01} - \hat{p}_{00} = 54.6 - 11.3 = 43.3$ excess deaths per 100,000 people, attributable to asbestos exposure by itself.

The effect on lung cancer death attributable to dual exposure appears to exceed the sum of the exposures' separate effects. The interaction contrast for the data example: $p_{11} - p_{10} - p_{01} + p_{00}$ indicates that the risk of lung cancer death in those who experience both exposures exceeds, by about 437.6 deaths per 100,000, the sum of the separate risks from smoking or from asbestos exposure. Section 3.3 illustrates a formal test of this departure from additivity of effects.

Calculated for the data example, the RERI, which quantifies additivity of effects on the relative risk scale, $RR_{11} - RR_{01} - RR_{10} + 1$ = [601.9/11.3] - [54.6/11.3] - [121.0/11.3] +1 = 38.7. Section 4.1 discusses approaches to formal testing of the RERI.

3.2 The linear binomial model, which assesses effects measured as probabilities and differences in probabilities (risk differences), detects a departure from additivity of effects in these data.

In 2.1, we introduced a definition of biological independence (Equation 1) and, in 2.3, linked it to Rothman's "interaction contrast" (Equation 4), which amounts to a formal test of additivity. We can estimate the interaction contrast in the linear binomial model described by Spiegelman and Hertzmark (2005) and by Richardson and colleagues (2015), who employ it as a final step in a marginal structural model. The linear binomial model estimates effects in terms of probabilities and differences in probabilities:

$$P(Y = 1) = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ \qquad \text{(Equation 6)}$$

Recalling that X and Z take values of 1 for "exposure" and 0 for "no exposure", then

$$\hat{p}_{00} = \beta_0$$
$$\hat{p}_{10} - \hat{p}_{00} = (\beta_0 + \beta_1) - \beta_0 = \beta_1$$
$$\hat{p}_{01} - \hat{p}_{00} = (\beta_0 + \beta_2) - \beta_0 = \beta_2$$
$$\hat{p}_{11} - \hat{p}_{00} = (\beta_0 + \beta_1 + \beta_2 + \beta_3) - \beta_0 = \beta_1 + \beta_2 + \beta_3$$

Substituting these expressions into Equation 1, which defines additivity of effects,

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

$$\beta_1 + \beta_2 + \beta_3 = \beta_1 + \beta_2$$

In the linear binomial model, effects are additive if β_3, the regression coefficient associated with the product or interaction term, is equal to zero.

Substituting the expressions into Equation 4, we see that the model's estimate for β_3 is itself an estimates of the interaction contrast:

$$p_{11} - p_{10} - p_{01} + p_{00} = (\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_1) - (\beta_0 + \beta_2) + \beta_0 = \beta_3$$

In the linear binomial model, estimation of the interaction contrast is equivalent to estimation of the X*Z interaction. Both provide direct tests of additivity. Evidence against the hypothesis that $\beta_3 = 0$ or, equivalently, that the interaction contrast is equal to zero, is evidence of a departure from additivity.

Appendix B illustrates the construction of the linear binomial model using SAS PROC GENMOD (Spiegelman and Herzmark, 2005; Richardson et al., 2015). The syntax includes a MODEL statement that identifies the independent variables smk (smoking status) and asbestos (asbestos exposure status) and includes smk*asbestos, the interaction between smoking status and asbestos exposure. Options in the MODEL statement specify that the outcome (lung cancer) follows a binomial distribution, and link the outcome directly (through an identity link, not through a log or logit transformation) to the linear predictor on the right side of Equation 6. An

ESTIMATE statement defines the interaction contrast's four coefficients $(1\ -1\ -1\ \ 1)$, estimates the magnitude of non-additivity, and tests the null hypothesis that non-additivity is equal to zero. Evidence against the null hypothesis suggests "non-additivity" or interaction.

The model's point estimates for the number of deaths per 100,000 workers, generated using an LSMEANS statement and presented in Table 3, and are equal to those reported in Table 2. Table 3 also reports the model's estimates (and 95% CI) for regression coefficients. Two of these coefficients estimate the effects on lung cancer mortality of smoking among those not exposed to asbestos (β_1) and of asbestos exposure in non-smokers (β_2). The model's estimates for β_1 and β_2 are equal to corresponding effects reported in 3.1.

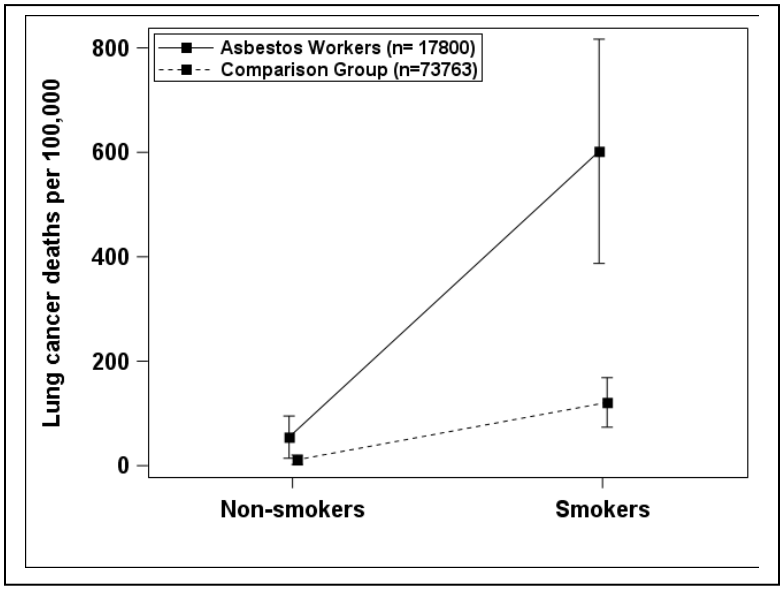
Table 3. Absolute risks (and risk differences) for death from lung cancer (per 100,000 workers) for those with exposure to asbestos and/or cigarette smoking, estimated by linear binomial model

	smk	asbestos	Estimate	Deaths per 100,000	95% CI on estimate	
					Lower	Upper
p_{11}	1 (yes)	1 (yes)	0.006019	601.926	387.183	816.669
p_{10}	1 (yes)	0 (no)	0.001210	121.048	73.627	168.469
p_{01}	0 (no)	1 (yes)	0.000546	54.619	14.169	95.070
p_{00}	0 (no)	0 (no)	0.000113	11.298	2.258	20.337
β_1	smk $(\hat{p}_{10} - \hat{p}_{00})$		0.001098	109.750	61.475	158.025
β_2	asbestos $(\hat{p}_{01} - \hat{p}_{00})$		0.000433	43.322	1.873	84.770
β_3	smk*asbestos		0.004376	437.557	213.768	661.345
IC	p_{11}-p_{10}-p_{01}+p_{00}		0.004376	437.557	213.768	661.345

The coefficient β_3, which is associated with the smk*asbestos interaction, is equivalent to the interaction contrast. The linear binomial model produces identical inference for β_3 and for the IC (estimate: 437.6 deaths per 100,000; 95% CI: 213.8, 661.3; p=0.00012702). Both estimates

match the calculation for the IC illustrated in Section 3.1. The consistency between the p values generated for the IC by the ESTIMATE statement, and for the smk*asbestos interaction by the MODEL statement, verifies that these two statistics offer equivalent tests of the null hypothesis that the effects of smoking and asbestos exposure are additive. Figure 1 illustrates the heterogeneity of the effects of smoking on lung cancer mortality in groups defined by asbestos exposure.

Figure 1. Biological interaction between asbestos exposure and smoking, illustrated as a non-additivity or heterogeneity of effects



Appendix B contains the SAS data and procedure steps that produced Table 3 and Figure 1.

3.3 Log binomial and logistic regression models, which assess effects measured as relative risks or odds ratios, do not detect a departure from multiplicativity of effects in these data.

In contrast to the linear binomial model, models that employ logarithmic transformations of probabilities (log binomial models) or their corresponding odds (logistic regression models) assess departures from multiplicativity of effects. Multiplicativity of effects is defined in a manner analogous to the definition of additivity of effects. The effects of two causal factors (X and Z) on an outcome (Y) are multiplicative if their joint effects are equal to the *product* of their separate and independent effects. Under this definition, when effects are multiplicative, relative risks will conform to the relationship: $\mathrm{RR_{XZ}} = \mathrm{RR_X} \times \mathrm{RR_Z}$, and odds ratios will conform to the relationship: $\mathrm{OR_{XZ}} = \mathrm{OR_X} \times \mathrm{OR_Z}$. A log binomial model

$$\ln[\mathrm{P}(\mathrm{Y}=1)] = \beta_0 + \beta_1 \mathrm{X} + \beta_2 \mathrm{Z} + \beta_3 \mathrm{XZ},$$

can estimate and test the multiplicativity of relative risks.

Appendix C shows in detail how, after restating the log binomial model as:

$$P(Y = 1) = \exp(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ).$$

it follows that: $RR_{XZ} = \exp(\beta_1 X + \beta_2 Z + \beta_3 XZ)$; $RR_X = \exp(\beta_1 X)$; $RR_Z = \exp(\beta_2 Z)$.

If there is no departure from multiplicativity among relative risks, then:

$$RR_{XZ} = RR_X \, RRz$$

$$\exp(\beta_1 X + \beta_2 Z + \beta_3 XZ) = \exp(\beta_1 X)\exp(\beta_2 Z) = \exp(\beta_1 X + \beta_2 Z).$$

This equality holds only if $\beta_3 = 0$, where β_3 is the regression coefficient associated with the product term. If the log binomial model's estimate of β_3, or its test of the null hypothesis $\beta_3 = 0$ suggests that $\beta_3 \neq 0$, that constitutes evidence of a departure from multiplicativity among the relative risks.

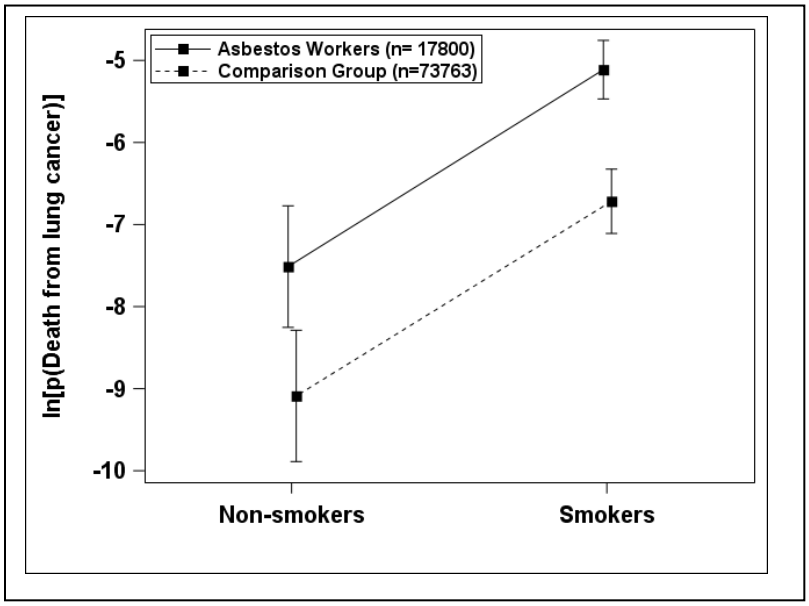
Similarly, the logistic regression model

$$\ln[P(Y = 1)/P(Y = 0)] = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

estimates and tests the multiplicativity of odds and odds ratios. Appendix C provides an extended explanation of how the inclusion of product terms in the log binomial and logistic regression models provides direct tests of the null hypothesis that there is no departure from multiplicativity of effects.

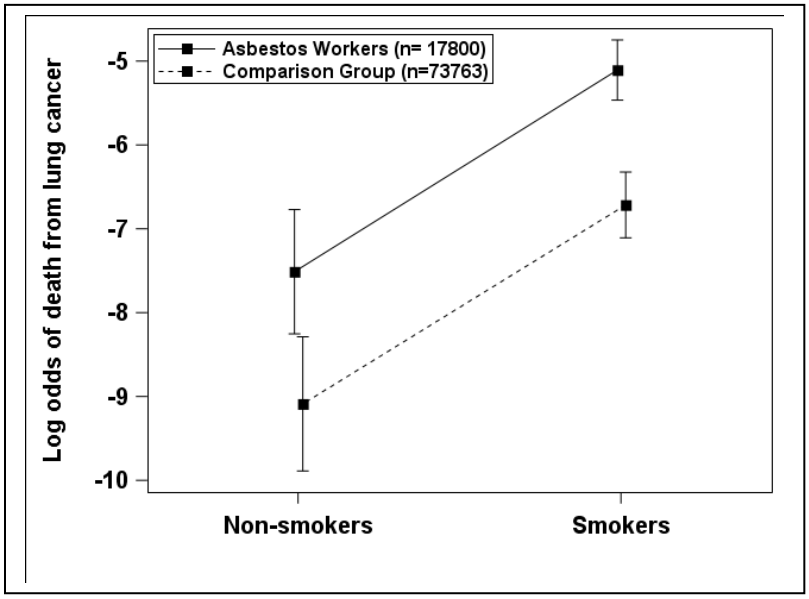
The log binomial finds no evidence of statistical interaction between smoking and asbestos exposure in the data example (p=0.9637). A depiction of the model's estimates (Figure 2) shows no heterogeneity nor departure of multiplicativity of effects. Appendix D shows the SAS data and procedure steps that generated the log binomial model and the depiction of its estimates in Figure 2.

Figure 2. Predicted log probabilities illustrate a lack of departure from multiplicativity of effects in the log binomial model.



Similarly, a logistic regression model, which employs a logit transformation of the outcome's probability, finds no evidence of departure from multiplicativity of effects. The model fails to reject the null hypothesis that there is no statistical interaction between smoking and asbestos exposure (p=0.9581). A depiction of the model's estimates (Figure 3) shows no heterogeneity of effects. Appendix E shows the SAS data and procedure steps that generated the logistic regression model and the depiction of its estimates in Figure 3.

Figure 3. Predicted log odds illustrate a lack of departure from multiplicativity of effects in the logistic regression model.



4. Choosing among statistical models

Although neither the log binomial nor the logistic regression model detects statistical interaction in the data example, this is not evidence of a lack of biological interaction. The finding underscores, however, the importance of constructing a statistical model that estimates an effect measure that is relevant to the scientific question. The three statistical models discussed here each estimate different effect measures, and each has advantages and limitations.

4.1 Choosing log binomial or logistic regression models that generate estimates of the RERI

Although the log binomial and logistic regression models test statistical interactions that relate to departures from a multiplicativity of effects, they are widely used in epidemiology to assess biological interaction that is hypothesized to manifest as a departure from additivity. The RERI, which assesses departures from additivity of effects on a relative risk or odds ratio scale (Vandenbroucke, von Elm, Altman, et al., 2007), is generated from these models' estimates.

However widespread, use of the RERI has disadvantages. Because it is constructed from ratios, the RERI cannot be interpreted as the number of excess deaths attributable to exposure to both smoking and asbestos. The RERI of 38.7, calculated for the data example, lacks the ease of interpretation of the IC's estimate, calculated in the linear binomial model, of 437.6 excess deaths per 100,000 (Table 3.)

A second disadvantage of relying on the RERI to assess departures from additivity involves the difficulty in obtaining standard errors with which to construct confidence intervals for its estimate, or to test hypotheses related to it. An influential approach, introduced by Hosmer and Lemeshow (1992), estimates the RERI using logistic regression and obtains standard errors for its estimates using the delta method. SAS syntax for their approach is provided by Andersson et al. (2005), and by Richardson and Kaufman (2009), who construct a "linear odds ratio model" using SAS PROC NLMIXED. Richardson and Kaufmann (2009) recommend bootstrapping as an alternative approach for obtaining confidence intervals. An empirical 95% confidence interval on the RERI, calculated for these data using 500 bootstrap samples, is 15.9, 132.6. However, the bounds for the RERI's confidence intervals present the same challenges to interpretation as the estimate itself.

4.2 Choosing the linear binomial model that directly estimates risks and risk differences

Logistic regression is widely used in epidemiology to study binomial outcomes in part because its use of the logit link, which is the canonical link for a binomial response, affords desirable statistical properties. Among these is logistic regression's reliability in converging on parameter estimates. Models that use other link functions can encounter problems with convergence. Zou

(2004) and Spiegelman and Herzmark (2005) point to the problem of convergence in the log binomial model and advocate use of a modified Poisson model to address the problem when it arises.

The linear binomial model, which uses the non-canonical identity link, can also fail to converge on solutions or estimates. This limitation interferes with the model's wider acceptance, despite its suitability for assessing additivity of effects through its direct estimation of probabilities, differences in probabilities and of the interaction contrast.

To address the issue of non-convergence, Spiegelman and Herzmark (2005) advocate modifying the linear binomial model so that it retains the identity link but assumes that the outcome follows a Poisson distribution. While this approach may ensure convergence, imposing the Poisson assumption causes the model to mis-specify the variance of a binomial outcome. This intentional mis-specification of the outcome's distribution reduces the efficiency of the model's standard errors, and of the hypothesis tests and confidence intervals that are based on those standard errors. Accordingly, Spiegelman and Herzmark (2005) recommend calculating standard errors that are robust despite misspecification. Appendix B contains the syntax for these modifications, including the incorporation of the REPEATED statement in SAS PROC GENMOD to initiate the GEE estimation of robust standard errors.

Richardson et al. (2015) do not suggest modifying the linear binomial model but, because they apply the model to weighted data, they also advocate the calculation of robust standard errors. Alternatively, the widespread availability of computing capacity makes attractive the bootstrapping of confidence intervals for parameter estimates, including the interaction contrast, as Richardson and Kaufmann (2009) have advocated for the RERI.

Cheung (2007) addresses non-convergence by proposing a modified least squares (MLS) model that, like the linear binomial model, uses an identity link. Like the approaches discussed above, Cheung's calculates robust standard errors. In contrast, Cheung's approach avoids specifying the outcome's assumed distribution by using ordinary least squares (OLS) instead of maximum likelihood estimation (MLE). This strategy cures the problem of non-convergence but does not guarantee that estimated probabilities will be in the logical range from 0 to 1.

5. Conclusions

Biological interaction is often hypothesized to manifest itself as a non-additivity of effects quantified as probabilities or absolute differences in probabilities. Applied to a data example that is widely used in epidemiology education to illustrate biological interaction, a linear binomial model detects statistical interaction while logistic and log binomial models do not.

The result affirms the consensus that biological interaction should generally be assessed as a departure from an additivity of effects. Statistics like the RERI, which are widely used in epidemiology, assess additivity on a relative risk scale. The linear binomial model produces estimates that, in contrast with the RERI and other ratio measures, are directly interpretable as probabilities and differences in probabilities.

Researchers can construct the linear binomial model, and obtain estimates and confidence intervals for the interaction contrast and other effects, using available software for generalized linear models. The model deserves wider use in research and judicious use in the teaching of epidemiology. The linear binomial model can encounter problems with convergence, but strategies exist to address this limitation.

Appendix

Appendix A. SAS program that created data example and produced Table 2

```
data one;
  array asbn (2) (73763 17800); /*n in published study*/
  array rate (4) (11.3 122.6 58.4  601.6);
                    /*lung ca deaths per 100k in published study*/
  smokeprev=0.28;
  do asbestos=0 to 1;
    do smk=0 to 1;
        do lungcadeath=0 to 1;
           mult=rate [2*asbestos + smk +1] / 100000;
           count1=asbn[asbestos+1] *
                   (abs((1-smk)-smokeprev)) *
               (abs((1-lungcadeath)-mult))  ;
           count=round(count1,1);
           output;
        end;
    end;
  end;
keep asbestos smk lungcadeath count;
run;
proc sort data=one (keep=asbestos smk lungcadeath count) out=two;
  by descending asbestos descending smk descending lungcadeath;
run;

/*version of dataset with individual observations*/
data long;
  set two;
  do i=1 to count;
    id+1;
    output;
  end;
run;

proc format;
  value smkf 1="Smokers" 0="Non-smokers";
  value gpf   1="Asbestos Workers (n= 17800)"
              0="Comparison Group (n=73763)";
  value death 1="Deaths due to lung CA"
              0="Alive or dead due to other causes";
run;

/*Table 2. Lung cancer deaths (per 100,000 workers) among those with exposure
to asbestos and/or cigarette smoking*/
proc freq data=two order=data;
  weight count;
  tables asbestos*smk*lungcadeath / nocol nopct outpct out=three;
  format smk smkf. asbestos gpf. lungcadeath death.;
run;

data four;
  set three;
  perhunthou=pct_row*1000;
```

```
run;

proc report nowd data=four;
  where lungcadeath=1;
  columns smk asbestos, perhunthou;
  define smk / group "Cigarette smoking" format=smkf. order=data;
  define asbestos / across "Asbestos Exposure" format=gpf. order=data;
  define perhunthou / analysis '' format=6.2;
run;
```

Appendix B. SAS steps that construct the linear binomial model and produce Table 3 and Figure 1

```
/*linear binomial model (after Spiegelman and Herzmark, 2005)   */
proc genmod data=long descending;
  class smk (ref=first) asbestos (ref=first)  ;
  model lungcadeath = smk asbestos smk*asbestos
         / link=identity dist=bin type3 wald ;
  lsmeans smk*asbestos / cl;
  ods output lsmeans=lsmeans estimates=estimates parameterestimates=betas
modelanova=type3;
  estimate "IC" smk*asbestos 1 -1 -1 1;
run;

/*linear binomial model that uses robust standard errors as advocated by
Richardson et al., 2015*/
proc genmod data=long descending;
  class smk (ref=first) asbestos (ref=first) id;
  model lungcadeath = smk asbestos smk*asbestos
         / link=identity dist=bin type3 wald ;
  repeated subject=id / type=ind;
  lsmeans smk*asbestos / cl;
  ods output lsmeans=lsmeans estimates=estimates parameterestimates=betas
modelanova=type3;
  estimate "IC" smk*asbestos 1 -1 -1 1;
run;

/*modification of linear binomial model advocated by Spiegelman and Herzmark
(2005) for instances when convergence fails*/
proc genmod data=long descending;
  class smk (ref=first) asbestos (ref=first) id;
  model lungcadeath = smk asbestos smk*asbestos
         / link=identity dist=poisson type3 wald ;
  repeated subject=id / type=ind;
  lsmeans smk*asbestos / cl;
  ods output lsmeans=lsmeans estimates=estimates parameterestimates=betas
modelanova=type3;
  estimate "IC" smk*asbestos 1 -1 -1 1;
run;


/*Table 3*/
/*Estimates for tabulated risks*/
data mortality;
  set lsmeans;
  mortality=estimate*100000;
  ucl=upper*100000;
  lcl=lower*100000;
run;
proc print noobs data=mortality;
  var smk asbestos estimate mortality lcl ucl;
run;

/*estimate for interaction contrast*/
data ic;
  set estimates;
```

```
  ic=meanestimate*100000;
  ic_lcl=meanlowercl*100000;
  ic_ucl=meanuppercl*100000;
run;
proc print noobs data=ic;
  var label meanestimate  ic ic_lcl ic_ucl probchisq;
  format meanestimate 9.6 probchisq 12.8 ;
run;

/*estimates of regression coefficients, showing that they
reflect excess deaths*/
data beta2;
  set betas (where=(df=1));
  excessdeaths=estimate*100000;
  ucl=upperwaldcl*100000;
  lcl=lowerwaldcl*100000;
run;
proc print noobs data=beta2;
  var parameter estimate excessdeaths lcl ucl probchisq;
  format estimate 9.6 probchisq 12.8;
run;

/*Figure 1. The linear binomial model's predicted
  Probabilities illustrate a departure from additivity.*/
proc template;
  define style styles.mystyle;
  parent=styles.default;
    class graphbackground / color=white;
    style GraphData1 from GraphData1 /
          contrastcolor=black linestyle=1;
    style GraphData2 from GraphData2 /
          contrastcolor=black linestyle=2;
  end;
run;
ods html style=styles.mystyle;
proc sgplot data=mortality;
  series y=mortality x=smk / group=asbestos name="one"
            groupdisplay=cluster clusterwidth=0.05
            markers markerattrs=(symbol=squarefilled size=10);
  highlow x=smk high=ucl low=lcl / group=asbestos
            groupdisplay=cluster clusterwidth=0.05
            type=line lineattrs=(pattern=1) lowcap=serif highcap=serif;
  xaxis values=(0 1) label=" " valueattrs=(size=14 weight=bold);
  yaxis label="Lung cancer deaths per 100,000"
        labelattrs=(size=14 weight=bold)
        valueattrs=(size=14 weight=bold);
  format smk smkf. asbestos gpf.;
  keylegend "one" / title="" location=inside down=2 position=topleft
             valueattrs=(size=12 weight=bold) ;
run;
ods html close;
```

Appendix C. Assessing departures from multiplicativity of effects in statistical models that use relative risks or odds ratios as their effect measures

We define multiplicativity of effects in a manner analogous to the way we defined additivity of effects. The effects of two exposures (X and Z) on an outcome (Y) are multiplicative if their joint effects are equal to the *product* of their separate and independent effects. When effects are multiplicative, we can predict that:

$$\frac{P(Y=1|X=1,Z=1)}{P(Y=1|X=0,Z=0)} = \frac{P(Y=1|X=1,Z=0)}{P(Y=1|X=0,Z=0)} \times \frac{P(Y=1|X=0,Z=1)}{P(Y=1|X=0,Z=0)},$$

or, equivalently, $RR_{XZ} = RR_X \times RR_Z$. We can similarly compare the joint and independent effects of two exposures using the odds ratio scale: $OR_{XZ} = OR \times OR_Z$.

These equations define null hypotheses that propose that *there is no departure from multiplicativity of effects.* Evidence against these equalities suggests a departure from multiplicativity of effects.

Log binomial models, which estimate relative risks as the measure of association, can test the null hypothesis that $RR_{XZ} = RR_X \times RR_Z$. Logistic regression models, which estimate odds ratios, can test the null hypothesis that $OR_{XZ} = OR_X \times OR_Z$. In either model, the inclusion of product terms provides direct tests of the null hypothesis that there is no departure from multiplicativity of effects.

Testing for departures from multiplicativity of effects in the log binomial model

In the log binomial model:

$$\ln[P(Y=1)] = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ,$$

or equivalently,

$$P(Y=1) = \exp(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ).$$

The model yields expressions for relative risks or, depending on the sampling scheme, a prevalence proportion ratio. These are ratios of two exponentiated linear functions:

$$RR_{XZ} = exp(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ)/exp(\beta_0) = exp(\beta_1 X + \beta_2 Z + \beta_3 XZ),$$

$$RR_X = exp(\beta_0 + \beta_1 X)/exp(\beta_0) = exp(\beta_1 X),$$

$$RR_Z = exp(\beta_0 + \beta_2 Z) \,/\, exp(\beta_0) = \; exp(\beta_2 Z),$$

Therefore, we have

$$RR_X RR_Z = exp(\beta_1 X)exp(\beta_2 Z) = exp(\beta_1 X + \beta_2 Z).$$

If there is no departure from multiplicativity among relative risks, then:

$RR_{XZ} = RR_X\ RRz$ and

$exp(\beta_1 X + \beta_2 Z + \beta_3 XZ) = exp(\beta_1 X + \beta_2 Z).$

Note that the equality holds if $\beta_3 = 0$, where β_3 is the regression coefficient associated with the product term. The log binomial model produces a test of the hypothesis $H_0: \beta_3 = 0$. Evidence that $\beta_3 \neq 0$ leads us to reject H_0 and suspect a departure from multiplicativity among the relative risks.

Testing for departures from multiplicativity of effects in the logistic regression model

Logistic regression models apply similar logic and similar algebra to assess multiplicativity among odds ratios. The logistic regression model is of the form:

$$\ln \frac{P(Y=1)}{P(Y=0)} = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ,$$

so that odds are expressed as

$$\frac{P(Y=1)}{P(Y=0)} = \exp(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ) = e^{\beta_0} e^{\beta_1 X} e^{\beta_2 Z} e^{\beta_3 XZ} \qquad \text{(A)}$$

With respect to odds, the logistic regression model's estimates are *multiplicative*. Its estimated coefficients relate (after exponentiation) to an "e^{β}*-fold* difference" in the odds of the outcome for a "one-unit" difference in a predictor variable. Similarly, coefficients estimated in a log binomial model predict, after exponentiation, the fold-difference in an outcome's probability associated with a "one-unit difference in a predictor variable.

If there is no departure from multiplicativity among odds ratios, then $OR_{XZ} = OR_X \times OR_Z$, which implies

$$\frac{\exp(\beta_0+\beta_1\times1+\beta_2\times1+\beta_3\times1\times1)}{\exp(\beta_0+\beta_1\times0+\beta_2\times0+\beta_3\times0\times0)} = \frac{\exp(\beta_0+\beta_1\times1+\beta_2\times0+\beta_3\times1\times0)}{\exp(\beta_0+\beta_1\times0+\beta_2\times0+\beta_3\times0\times0)} \times \frac{\exp(\beta_0+\beta_1\times0+\beta_2\times1+\beta_3\times0\times1)}{\exp(\beta_0+\beta_1\times0+\beta_2\times0+\beta_3\times0\times0)}.$$

This equality simplifies to

$$\exp(\beta_1 + \beta_2 + \beta_3) = \exp(\beta_1 + \beta_2).$$

The equality holds only if $\beta_3 = 0$, where β_3 is the regression coefficient associated with the product term. The logistic regression model produces a test of the hypothesis $H_0: \beta_3 = 0$. Evidence that $\beta_3 \neq 0$ leads us to reject H_0 and suspect a departure from multiplicativity among the odds ratios.

Appendix D. SAS data and procedure steps that generated the log binomial model and the depiction of its estimates in Figure 2.

```
proc genmod data=long descending;
  class smk (ref=first) asbestos (ref=first)  ;
  model lungcadeath = smk asbestos smk*asbestos
         / link=log dist=bin type3 wald lrci;
  lsmeans smk*asbestos / cl;
  ods output lsmeans=lsmeans ;
run;

ods html style=styles.mystyle;
proc sgplot data=lsmeans;
  series y=estimate x=smk / group=asbestos name="one"
           groupdisplay=cluster clusterwidth=0.05
           markers markerattrs=(symbol=squarefilled size=10);
  highlow x=smk high=upper low=lower / group=asbestos
           groupdisplay=cluster clusterwidth=0.05
           type=line lineattrs=(pattern=1) lowcap=serif highcap=serif;
  xaxis values=(0 1) label=" " valueattrs=(size=14 weight=bold);
  yaxis label="ln[p(Death from lung cancer)]"
        labelattrs=(size=14 weight=bold)
        valueattrs=(size=14 weight=bold);
  format smk smkf. asbestos gpf.;
  keylegend "one" / title="" location=inside down=2 position=topleft
            valueattrs=(size=12 weight=bold) ;
run;
ods html close;
```

Appendix E. SAS data and procedure steps that generated the logistic model and the depiction of its estimates in Figure 3.

```
proc genmod data=long descending;
  class smk (ref=first) asbestos (ref=first)  ;
  model lungcadeath = smk asbestos smk*asbestos
          / link=logit dist=bin type3 wald lrci;
  lsmeans smk*asbestos / cl;
  ods output lsmeans=lsmeans ;
run;

ods html style=styles.mystyle;
proc sgplot data=lsmeans;
  series y=estimate x=smk / group=asbestos name="one"
            groupdisplay=cluster clusterwidth=0.05
            markers markerattrs=(symbol=squarefilled size=10);
  highlow x=smk high=upper low=lower / group=asbestos
            groupdisplay=cluster clusterwidth=0.05
            type=line lineattrs=(pattern=1) lowcap=serif highcap=serif;
  xaxis values=(0 1) label=" " valueattrs=(size=14 weight=bold);
  yaxis label="Log odds of death from lung cancer"
        labelattrs=(size=14 weight=bold)
        valueattrs=(size=14 weight=bold);
  format smk smkf. asbestos gpf.;
  keylegend "one" / title="" location=inside down=2 position=topleft
             valueattrs=(size=12 weight=bold) ;
run;
ods html close;
```

Acknowledgement:

Dr. Zhao's work was partially supported by funding provided by National Institutes of Health, National Institute of General Medical Sciences [Grant 1 U54GM104938, PI Judith James].

The authors thank Dr. Tabitha Garwe for important comments on the manuscript.

References:

Andersson, T., Alfredsson, L., Källberg, H., Zdravkovic, S., and Ahlbom, A. (2005), Calculating Measures of Biological Interaction, *European Journal of Epidemiology*, 20(7):575-579. doi:10.1007/s10654-005-7835-x

Bieler, G.S., Brown, G.G., Williams, R.L., and Brogan, D.J. (2010), Estimating Model-Adjusted Risks, Risk Differences, and Risk Ratios from Complex Survey Data, *American Journal of Epidemiology*, 171(5):618-623.

Cheung, Y.B. (2007), A Modified Least-Squares Regression Approach to the Estimation of Risk Difference, *American Journal of Epidemiology*, 166(11):1337-44.

Hammond, E.C., Selikoff, I.J., and Seidman, H. (1979), Asbestos Exposure, Cigarette Smoking and Death Rates, *Annals of the New York Academy of Science*, 330,473-90. http://onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.1979.tb18749.x/epdf

Hosmer, D.W., and Lemeshow, S. (1992), Confidence Interval Estimation of Interaction, *Epidemiology*, 3(5),452-456.

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*, 16,6,756–760. https://doi.org/10.1093/aje/kwn411

Richardson, D.B., Kinlaw, A.C., MacLehose, R.F., and Cole, S.R. (2015), Standardized Binomial Models for Risk or Prevalence Ratios and Differences. *International Journal of Epidemiology,* 44(5),1660-72. doi: 10.1093/ije/dyv137. Epub 2015 Jul 30.

Rothman, K. J. (1986), *Modern Epidemiology*, 1st Edition, Boston, MA: Little, Brown and Company.

Rothman, K.J. (2002), *Epidemiology: An Introduction*, New York, NY: Oxford University Press.

Rothman, K.J., and Greenland, S. (ed.) (1998), *Modern Epidemiology*, Philadelphia, PA: Lippincott Williams and Wilkins.

Spiegelman, D., and Hertzmark, E. (2005), Easy SAS Calculations for Risk or Prevalence Ratios and Differences, *American Journal of Epidemiology*, 162(3):199-200.

Szklo, M., and Nieto, F.J. (2004), *Epidemiology: Beyond the Basics*, 3rd Edition, Sudbury, MA: Jones and Bartlett.

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(6):805-835.

VanderWeele, T. J. (2013), Reconsidering the Denominator of the Attributable Proportion for Interaction. *European Journal of Epidemiology*, *28*(10), 779–784.

VanderWeele, T.J., and Knol, M.J. (2014), A Tutorial on Interaction, *Epidemiologic Methods*, 3,33-72.

Zou, G. (2004), A Modified Poisson Regression Approach to Prospective Studies With Binary Data, *American Journal of Epidemiology*, 159(7),702-706.

Contact information

David M. Thompson, Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104 (e-mail: dave-thompson@ouhsc.edu).

Yan Daniel Zhao, Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104 (e-mail: Daniel-zhao@ouhsc.edu).