

METHODOLOGY

Estimating interaction on an additive scale between continuous determinants in a logistic regression model

Mirjam J Knol,^{1,2*} Ingeborg van der Tweel,³ Diederick E Grobbee,¹ Mattijs E Numans¹ and Mirjam I Geerlings¹

Accepted 3 July 2007

Background To determine the presence of interaction in epidemiologic research, typically a product term is added to the regression model. In linear regression, the regression coefficient of the product term reflects interaction as departure from additivity. However, in logistic regression it refers to interaction as departure from multiplicativity. Rothman has argued that interaction estimated as departure from additivity better reflects biologic interaction. So far, literature on estimating interaction on an additive scale using logistic regression only focused on dichotomous determinants. The objective of the present study was to provide the methods to estimate interaction between continuous determinants and to illustrate these methods with a clinical example.

Methods and results From the existing literature we derived the formulas to quantify interaction as departure from additivity between one continuous and one dichotomous determinant and between two continuous determinants using logistic regression. Bootstrapping was used to calculate the corresponding confidence intervals. To illustrate the theory with an empirical example, data from the Utrecht Health Project were used, with age and body mass index as risk factors for elevated diastolic blood pressure.

Conclusions The methods and formulas presented in this article are intended to assist epidemiologists to calculate interaction on an additive scale between two variables on a certain outcome. The proposed methods are included in a spreadsheet which is freely available at: <http://www.juliuscenter.nl/additive-interaction.xls>.

Keywords Interaction, biologic, additive, continuous, effect-modification, relative excess risk

Background

In epidemiology, interaction refers to the situation where the effect of one risk factor (A) on a certain disease outcome is different across strata of another risk factor (B),

or vice versa. This means that if interaction between A and B is present, A and B are not independent in causing a certain disease. If the combined effect of A and B is larger (or smaller) than the sum of the individual effects of A and B, there is interaction on an additive scale or departure from additivity. Interaction on a multiplicative scale, or departure from multiplicativity, occurs when the combined effect of A and B is larger (or smaller) than the product of the individual effects.

Rothman discerns two types of interactions: statistical and biologic.¹ Statistical interaction means departure from the underlying form of a statistical model. Because there are various statistical models, statistical interaction does not have a

1 Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, The Netherlands.

2 Department of Pharmaco-Epidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences, Utrecht University, The Netherlands.

3 Centre for Biostatistics, Utrecht University, The Netherlands.

* Corresponding author. Str. 6.131, Julius Center, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands.
E-mail: m.j.knol@umcutrecht.nl

consistent meaning. Most researchers assess interaction by entering a product term into the linear or logistic regression model. However, the interpretation of the regression coefficient of the product term depends on the statistical model. In linear regression analysis the regression coefficient of the product term means departure from additivity, whereas in logistic regression (and in Cox regression) the regression coefficient of the product term estimates departure from multiplicativity (Appendix 1). Biologic interaction means that two causes are both needed to cause disease; the two causes are component causes in the same causal model. Rothman has argued that when biologic interaction is examined, we should focus on interaction as departure from additivity rather than departure from multiplicativity.¹ In aetiologic epidemiologic research, we are interested in biologic interaction rather than in statistical interaction. However, by adding a product term to a logistic model, interaction is (unknowingly) estimated as departure from multiplicativity.

Rothman² and Hosmer and Lemeshow³ have shown how interaction as departure from additivity can be quantified in a logistic regression model. They proposed to make one categorical variable with four levels that combines two dichotomous determinants. Assmann *et al.*⁴ demonstrated that bootstrapping may give better coverage of the 95% CI of the estimate of interaction than the delta method described by Hosmer and Lemeshow.³ These studies only focused on interaction between two dichotomous determinants. In epidemiologic research, we are often also interested in the effect of continuous determinants on an outcome and dichotomizing a continuous variable may lead to loss of information.⁵

This article illustrates with a clinical example the methods and formulas to estimate interaction, as departure from additivity between continuous determinants, and its uncertainty on both a dichotomous and a continuous outcome. We will first describe the concept of interaction on an additive scale and show the difference with interaction on a multiplicative scale using a simple 2×2 table. Then we will explain how logistic regression analysis can be used to estimate interaction on an additive scale. There are two reasons why regression analysis rather than simple tabulation might be needed to estimate interaction. First, there may be a need to adjust for confounders. Second, if one or both of the determinants are continuous it is not possible to construct a 2×2 table. In this article we will focus on one of the three measures for interaction as departure from additivity, namely the Relative Excess Risk due to Interaction (RERI). The other two measures, Attributable Proportion (AP) and Synergy index (S), will be shortly discussed in the application section of the article.

Example dataset

To illustrate the methods and formulas, we will use data from the Utrecht Health Project (UHP), which is described in detail elsewhere.⁶ In brief, the UHP is an ongoing longitudinal study, which started in the year 2001, among all inhabitants of a new residential area in the city of Utrecht, The Netherlands. At baseline, an Individual Health Profile is made, which is based on a questionnaire, physical examination and blood measurements. By January 2005, 13 128 subjects were

Table 1 Descriptive statistics of age, body mass index and diastolic blood pressure

	<i>N</i>	Mean	SD	Median	Min	Max
Age (years)	4897	39.3	12.5	35.6	18.0	91.1
Body mass index (kg/m ²)	4897	25.5	4.2	24.9	9.2	46.7
Diastolic blood pressure (mmHg)	4897	77.8	10.5	77.0	48.5	126.5

invited of whom 6755 gave informed consent (51.4%) and entry data were complete on 6304 (48.0%) adults and children. The adult UHP population consists of 2221 (44.9%) males and 2729 (55.1%) females with a mean age (SD) of 39.3 (12.5) years.

The data set we use in this article comprises two continuous determinants, age and body mass index (BMI), and one continuous outcome, diastolic blood pressure. Fifty-three subjects had a missing value on BMI or diastolic blood pressure and were excluded from the analyses. Table 1 presents descriptive statistics of the three variables. The three variables were dichotomized because our aim was to show how to calculate additive interaction using both dichotomous and continuous determinants and outcomes. We reasoned that using the same determinants and outcome throughout all the examples would enhance the understanding of the article. As a result, we did not use truly dichotomous determinants, such as gender. Age was dichotomized according to an arbitrarily chosen cutpoint of 40 years, where age <40 years was coded as 0 and age ≥40 years was coded as 1. BMI was dichotomized according to the overweight cutpoint of 25 kg/m², where BMI <25 kg/m² was coded as 0 and BMI ≥25 kg/m² as 1. Diastolic blood pressure was dichotomized according to a cutpoint of 90 mm Hg, where a normal blood pressure was coded as 0 and hypertension as 1. The percentages in the resulting categories were 65.2 and 34.8% for age, 50.9 and 49.1% for BMI and 87.5 and 12.5% for hypertension.

Estimating interaction on an additive scale using a 2×2 table

Consider age (*A*) and BMI (*B*) as dichotomous risk factors for diastolic hypertension (*D*). A 2×2 table can be constructed with the absolute risk of disease in the four following categories: young subjects with normal BMI (*A*–*B*–), older subjects with normal BMI (*A*+*B*–), young subjects with overweight (*A*–*B*+) and older subjects who are overweight (*A*+*B*+) . The risk in category *A*–*B*– is called the background risk because in this category the disease frequency is caused by other factors than *A* and *B*. Table 2 shows the absolute risks of hypertension in these categories and the risk differences and relative risks in strata of age and BMI.

Interaction on an additive scale is present if the combined effect of *A* and *B* is not equal to the sum of the effects of *A* and *B*:

$$(R_{A+B+} - R_{A-B-}) \neq (R_{A+B-} - R_{A-B-}) + (R_{A-B+} - R_{A-B-}) \quad (1)$$

where *R* indicates the absolute risk of disease in that specific stratum. Note that the background risk is subtracted to get the

Table 2 Absolute risks, risk differences and relative risks of diastolic hypertension according to strata of age and BMI

		Normal BMI (B-)		Overweight (B+)		Risk difference	Relative risk
		N	Risk	N	Risk		
Young (A-)	Normal tension (D-)	1731		1232			
	Hypertension (D+)	79	4.4%	153	11.0%	6.6%	2.50
Old (A+)	Normal tension (D-)	581		743			
	Hypertension (D+)	100	14.7%	278	27.2%	12.5%	1.85
Risk difference			10.3%		16.2%		
Relative risk			3.34		2.47		

effects of *A* alone, *B* alone and *A* and *B* combined. In our example: $(27.2 - 4.4) \neq (14.7 - 4.4) + (11.0 - 4.4) \Rightarrow 22.8 > 16.9$, meaning there is 'positive' interaction on an additive scale because the combined effect is larger than the sum of the individual effects. By dividing all risks in Formula (1) by the background risk, R_{A-B-} , an equivalent expression for risk ratios (RRs) is obtained:

$$(RR_{A+B+} - 1) \neq (RR_{A+B-} - 1) + (RR_{A-B+} - 1). \quad (2)$$

In our example: $(27.2/4.4 - 1) \neq (14.7/4.4 - 1) + (11.0/4.4 - 1) \Rightarrow 5.18 > 3.84$.

Interaction on a multiplicative scale is present if the combined effect of *A* and *B* is not equal to the product of the effects of *A* and *B*:

$$(R_{A+B+}/R_{A-B-}) \neq (R_{A+B-}/R_{A-B-}) * (R_{A-B+}/R_{A-B-}). \quad (3)$$

In our example: $(27.2/4.4) \neq (14.7/4.4) \times (11.0/4.4) \Rightarrow 6.18 < 8.35$, meaning that there is 'negative' interaction on a multiplicative scale because the combined effect is smaller than the product of the individual effects. The fact that interaction is present can also be seen in Table 2, as the effect of BMI on hypertension is different across strata of age. Also, the effect of age on hypertension is different across strata of BMI. However, the risk difference is highest in the older age stratum and in the overweight stratum, whereas the RR is highest in the younger age stratum and in the normal BMI stratum. This agrees with the calculations above, as these showed there is positive interaction on an additive scale and negative interaction on a multiplicative scale. This example illustrates that it depends on the measure of effect (risk difference or RR) whether interaction is present or not, or in which direction the interaction operates.

The amount of interaction as departure from additivity can be derived from Formula (1) for absolute risks:

$$(R_{A+B+} - R_{A-B-}) - (R_{A+B-} - R_{A-B-}) - (R_{A-B+} - R_{A-B-}) \quad (4)$$

and from Formula (2) for relative risks:

$$(RR_{A+B+} - 1) - (RR_{A+B-} - 1) - (RR_{A-B+} - 1). \quad (5)$$

Rothman called this amount of interaction RERI.² Rewriting Formula (5) gives:

$$RERI = RR_{A+B+} - RR_{A+B-} - RR_{A-B+} + 1. \quad (6)$$

In our example the RERI is $RERI = 6.18 - 3.34 - 2.50 + 1 = 1.34$, meaning that the relative risk for hypertension in older overweight subjects is 1.34 more than if there were no

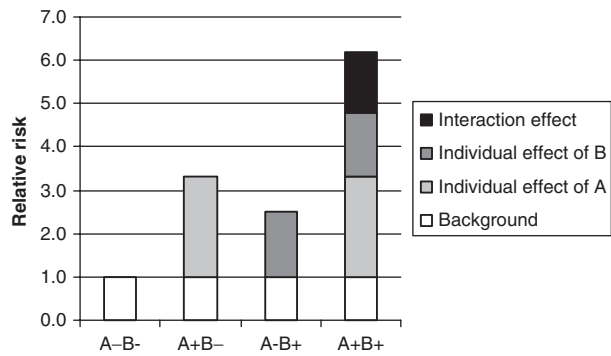


Figure 1 Relative risks in categories A-B-, A+B-, A-B+ and A+B+, divided in background, individual effect of *A*, individual effect of *B* and the interaction effect as departure from additivity

interaction between age and BMI. Figure 1 shows this graphically.

Note that the absolute background risk was 4.4% and thus the absolute risk due to interaction is 5.9% ($4.4\% \times 1.34$), which is exactly the amount of interaction for absolute risks calculated with Formula (4).

Assuming that the odds ratio (OR) approximates the relative risk, this formula can also be used for ORs. Note that in the absence of interaction as departure from additivity, i.e. when there is exact additivity, $RERI = 0$.

Estimating interaction on an additive scale using logistic regression

Determinant *A*, determinant *B* and the product of *A* and *B* are included in the logistic regression model. This may seem confusing because we indicated above that a product term tests departure from multiplicativity rather than additivity in a logistic regression model. However, as shown below, the regression coefficient of the product term can also be used to calculate interaction as departure from additivity. Including determinant *A*, *B* and the product of *A* and *B* in the logistic regression formula results in:

$$\ln\left(\frac{p}{1-p}\right) = \ln(\text{odds}) = \hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 B + \hat{\beta}_3 AB. \quad (7)$$

To calculate RERI the following three ORs are needed: (i) *A*+*B*- relative to *A*-*B*- which is $e^{\hat{\beta}_1}$, (ii) *A*-*B*+ relative

to A–B– which is $e^{\hat{\beta}_2}$ and (iii) A+B+ relative to A–B– which is $e^{\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3}$ (Appendix 1). The formulas to assess presence of interaction on an additive scale and to estimate the RERI are:

$$(e^{\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3} - 1) \neq (e^{\hat{\beta}_1} - 1) + (e^{\hat{\beta}_2} - 1) \tag{8}$$

and

$$\text{RERI} = e^{\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3} - e^{\hat{\beta}_1} - e^{\hat{\beta}_2} + 1. \tag{9}$$

The regression coefficients from the logistic regression model can be substituted into Formulae (8) and (9). Important to note is that these formulae can be used for two dichotomous as well as two continuous determinants or a combination of both.

Estimating confidence interval

A simulation study⁴ showed that the first bootstrap percentile method gave better coverage of the 95% CI than the delta method.³ Moreover, note that assessment of a continuous determinant per, e.g., 5 units instead of per 1 unit leads to a non-linear transformation of the RERI and its CI. Then the delta method cannot be used whereas bootstrapping can. For these two reasons, we adopted the first bootstrap percentile method to calculate the CI around the estimate of interaction. From the original data set 10 000 bootstrap samples (with replacement) were taken, each of which was the same size as the original sample. The RERI was then estimated in each of these new samples and the 95% CI for RERI was estimated as the 2.5th and 97.5th percentiles of the resulting bootstrap sampling distribution. The statistical program S-PLUS 6.2 was used (S-PLUS 6.2, Insightful, Seattle, USA) to carry out the bootstrapping procedure. Appendix 2 presents the script we used and an example output.

Empirical examples of estimating interaction on an additive scale

Two dichotomous determinants and dichotomous outcome

Consider age and BMI as risk factors for diastolic hypertension, all dichotomized as described before. Age, BMI and the product of age and BMI are entered as the independent variables and diastolic hypertension as the dependent variable in a logistic regression model. The output of the logistic regression model shows that an older person has a 3.77 times higher risk of diastolic hypertension than a young person (Table 3).

Table 3 Output of logistic regression model with age and BMI as dichotomous (dich) determinants and product of age and BMI entered into the model. Outcome is diastolic hypertension

Parameter	Estimate	Standard error	95% CI of OR		
			OR	Lower	Upper
Age dich	1.33	0.16	3.77	2.77	5.14
BMI dich	1.00	0.14	2.72	2.05	3.61
Age dich × BMI dich	-0.23	0.19	0.80	0.55	1.17
Constant	-3.09	0.12			

Overweight subjects have a 2.72 times higher risk of diastolic hypertension compared with subjects with normal BMI. The OR and CI of the product term provide some evidence of ‘negative’ interaction on a multiplicative scale (OR (95% CI)=0.80 (0.55–1.17)). Most readers will conclude that there is no interaction between age and BMI, because the OR of the product term is not significant. However, since the underlying model is a logistic regression model, this product term refers only to interaction on a multiplicative scale. The OR of 0.80 actually means that the combined effect of old age and overweight is 0.80 times the product of the individual effects of old age and overweight. To assess the amount of interaction on an additive scale we use Formula (8): $(e^{\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3} - 1) \neq (e^{\hat{\beta}_1} - 1) + (e^{\hat{\beta}_2} - 1) \Rightarrow 7.2 > 4.5$ and Formula (9): $\text{RERI} = e^{\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3} - e^{\hat{\beta}_1} - e^{\hat{\beta}_2} + 1 = 2.7$. There is ‘positive’ interaction on an additive scale because 7.2 is larger than 4.5. The amount of additive interaction, the RERI, is equal to 2.7. This means that the relative risk of having hypertension in older subjects with overweight is 2.7 more than if there were no interaction between age and BMI. Bootstrapping resulted in the following RERI with 95% CI: RERI (95% CI) = 2.7 (1.3, 4.4). Note that a RERI of 0 indicates exact additivity and thus no interaction on an additive scale. In this example, there is more evidence for interaction on an additive scale than for interaction on a multiplicative scale.

One continuous and one dichotomous determinant and dichotomous outcome

Exactly the same methods as described above can be used for one dichotomous and one continuous determinant. Consider again age and BMI as risk factors for hypertension but now age is a continuous variable. Age, BMI and the product of age and BMI are entered as the independent variables and hypertension as the dependent variable in a logistic regression model. We arbitrarily chose to evaluate the effect of age per 5 years increase by dividing age by 5, and included this variable in the model. Table 4 shows that per 5 years increase of age, the risk of having diastolic hypertension increases with a factor of 1.29. Overweight subjects have a 4.05 times higher risk of diastolic hypertension compared with subjects with normal BMI. The OR and CI of the product term provide evidence for ‘negative’ interaction on a multiplicative scale (OR (95% CI)=0.94 (0.88–1.00)). However, ‘positive’ interaction is present on an additive scale: $(e^{\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3} - 1) \neq (e^{\hat{\beta}_1} - 1) + (e^{\hat{\beta}_2} - 1) = 4.0 > 3.4$ and $\text{RERI} = e^{\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3} - e^{\hat{\beta}_1} - e^{\hat{\beta}_2} + 1 = 0.6$ with a 95% CI of (0.3; 1.0). A RERI of 0.6 means that the relative risk of having hypertension in overweight subjects is 0.6 more with each

Table 4 Output of logistic regression model with age (per 5 years) as continuous (cont) determinant, BMI as dichotomous (dich) determinant and product of age and BMI entered into the model. Outcome is diastolic hypertension

Parameter	Estimate	Standard error	95% CI of OR		
			OR	Lower	Upper
Age (per 5 years) cont	0.26	0.03	1.29	1.22	1.36
BMI dich	1.40	0.31	4.05	2.20	7.45
Age cont × BMI dich	-0.06	0.03	0.94	0.88	1.00
Constant	-4.60	0.25			

5 years of increase in age than if there were no interaction between age and BMI.

Two continuous determinants and dichotomous outcome

Consider again age and BMI as risk factors for diastolic hypertension but now as two continuous variables. Age, BMI and the product of age and BMI are entered as the independent variables and diastolic hypertension as the dependent variable in a logistic regression model. We again chose to evaluate the effect of age per 5 years. The effect of BMI was estimated per 2 units. Table 5 shows that per 5 years increase of age and per 2 units increase of BMI, the risk of having diastolic hypertension increased with a factor of 1.41 and 1.39, respectively. The OR and CI of the product term show that there is little evidence for interaction on a multiplicative scale (OR (95% CI)=0.99 (0.98–1.01)). Calculating interaction on an additive scale gives: $(e^{\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3} - 1) \neq (e^{\hat{\beta}_1} - 1) + (e^{\hat{\beta}_2} - 1) \Rightarrow 0.93 > 0.80$ and $RERI = e^{\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3} - e^{\hat{\beta}_1} - e^{\hat{\beta}_2} + 1 = 0.14$ with a 95% CI of (0.04; 0.34). A RERI of 0.14 means that with every 5 years increase in age, and 2 units increase in BMI, the relative risk of having hypertension is 0.14 more than if there were no interaction.

Robustness of RERI and confidence interval

In the examples above, the effect of age was assessed per 5 years increase of age and the effect of BMI per 2 units increase of BMI. To assess the robustness of RERI and CI, we calculated the RERI and its 95% confidence interval using different years of increase in age (1, 2, 5 and 10) and different units of increase in BMI (1, 2 and 5). The results are presented in Table 6. Note that the increase in RERI with increasing units is not linear with the

Table 5 Output of logistic regression model with age (per 5 years) as continuous (cont) determinant, BMI (per 2 kg/m²) as continuous (cont) determinant and product of age and BMI entered into the model. Outcome is diastolic hypertension

95% CI of OR					
Parameter	Estimate	Standard error	OR	Lower	Upper
Age (per 5 years) cont	0.34	0.10	1.41	1.15	1.72
BMI (per 2 kg/m ²) cont	0.33	0.07	1.39	1.22	1.59
Age cont × BMI cont	-0.01	0.01	0.99	0.98	1.01
Constant	-8.09	0.92			

Table 6 RERI and 95% confidence interval for different units increase in age (1, 2, 5 and 10) and BMI (1, 2 and 5)

RERI (95% CI)	Age			
	1	2	5	10
BMI				
1	0.011 (0.004–0.026)	0.023 (0.007–0.051)	0.064 (0.020–0.158)	0.152 (0.040–0.390)
2	0.025 (0.007–0.055)	0.051 (0.015–0.114)	0.139 (0.040–0.337)	0.327 (0.079–0.812)
5	0.078 (0.023–0.193)	0.161 (0.047–0.412)	0.440 (0.130–1.210)	1.031 (0.287–3.106)

units increase. For example, the RERI with age per 1 unit increase and BMI per 2 units increase (0.025, Table 6) is not exactly a factor 2 larger than the RERI with age per 1 unit increase and BMI per 1 unit increase (0.011, Table 6).

Estimating interaction on an additive scale using linear regression

As explained previously, in case of linear regression the regression coefficient of the product term reflects interaction as departure from additivity (Appendix 1). Consider again age and BMI as risk factors for elevated diastolic blood pressure. Age and BMI are dichotomous variables and diastolic blood pressure is a continuous variable. Age, BMI and the product of age and BMI are entered as the independent variables and diastolic blood pressure as the dependent variable in a linear regression model. The output is presented in Table 7. The combined effect of age and BMI is larger than the sum of the individual effects of age and BMI: $(84.2 - 73.7) \neq (79.1 - 73.7) + (77.7 - 73.7) \Rightarrow 10.5 > 9.4$ or $(5.4 + 4.0 + 1.1 - 0) > (5.4 - 0) + (4.0 - 0) \Rightarrow 10.5 > 9.4$.

The amount additive interaction is: $(5.4 + 4.0 + 1.1) - 5.4 - 4.0 + 0 = 1.1$, which (by definition) equals the regression coefficient of the product term. Note that this estimate of interaction is not the same as ‘RERI’ as this calculation concerns the change in absolute values of the continuous outcome instead of a change in the relative risks. The confidence interval around the interaction estimate is easily calculated with the standard error of the regression coefficient of the product term: (-0.1; 2.3). So there is considerable evidence for ‘positive’ interaction on an additive scale between age and BMI as dichotomous risk factors for diastolic blood pressure on a continuous scale.

Practical use

Besides the RERI, Rothman has proposed two other measures of interaction on an additive scale: the proportion of disease among those with both exposures that is attributable to their interaction (AP) and the ratio between the combined effect and the sum of the individual effects, the synergy index (S)². The formulas of these measures are:

$$AP = \frac{RERI}{RR_{A+B+}}, \tag{10}$$

and

$$S = \frac{RR_{A+B+} - 1}{(RR_{A+B-} - 1) + (RR_{A-B+} - 1)}. \tag{11}$$

Table 7 Output of linear regression model with age and BMI as dichotomous determinants and product of age and BMI entered into the model. Outcome is diastolic blood pressure

95% CI of β					
Parameter	Estimate	Standard error	Lower	Upper	
Age dich	5.4	0.4	4.5	6.2	
BMI dich	4.0	0.3	3.4	4.7	
Age dich × BMI dich	1.1	0.6	-0.1	2.3	
Constant	73.7	0.2	73.2	74.1	

Note that in the absence of interaction as departure from additivity, AP is 0 and S is 1. These measures of interaction can also be calculated in case of one or two continuous determinants using the same approach as described above for the RERI. The CI can also be obtained by bootstrapping.

Assuming that a hazard ratio approximates a relative risk, the methods to estimate interaction on an additive scale described in this article can also be applied to Cox regression. The script for bootstrapping, however, should be adapted.

The proposed methods to estimate interaction on an additive scale with continuous determinants are included in a spreadsheet which is freely available at: www.juliuscenter.nl/additive-interaction.xls. In the spreadsheet the output of the logistic regression model (or Cox regression model) has to be filled in and all estimates of interaction (RERI, AP and S) are calculated. Furthermore, the script for bootstrapping in S-PLUS to calculate the 95% CI is presented in Appendix 2 and included in the spreadsheet.

In this article, we took Rothman's theory about the causal pie model as a starting point. This theory implies that biologic interaction is present if two causes are both needed to cause disease and therefore should be assessed as departure from additivity rather than multiplicativity. Not all researchers may agree with this view and the relevance of interaction on a multiplicative scale may be different in non-aetiologic research. However, when interaction on an additive scale is the measure of interest, the methods outlined in this article may be used fruitfully.

Conclusions

The aim of our article was to show that interaction as departure from additivity between continuous determinants can be estimated using logistic regression analysis and to give an empirical example. The methods and formulas presented in this article are intended to assist epidemiologists to calculate interaction on an additive scale between continuous determinants. To facilitate its use, the proposed methods are included in a spreadsheet which is freely available at: www.juliuscenter.nl/additive-interaction.xls.

Acknowledgements

We acknowledge the participating inhabitants of Leidsche Rijn, Utrecht, The Netherlands, and the general practitioners working in this area for providing research data from routine care. This study was supported by an unrestricted grant from Novo Nordisk and the Scientific Institute of Dutch Pharmacists (WINAp) and was supported by The Netherlands Organization for Scientific Research (ZON-MW 917.66.311). The Utrecht Health Project (LRGP) received grants from the Ministry of Health, Welfare, and Sports (VWS), the University of Utrecht, the Province of Utrecht, the Dutch Organisation of Care Research (ZON), the University Medical Center Utrecht (UMC Utrecht) and the Dutch College of Healthcare Insurance Companies (CVZ).

Conflict of interest: None declared.

KEY MESSAGES

- In linear regression the regression coefficient of the product term reflects interaction as departure from additivity, whereas in logistic regression it refers to interaction as departure from multiplicativity.
- So far, literature on estimating interaction as departure from additivity using logistic regression only focused on dichotomous determinants.
- The aim of our article is to show with an empirical example that additive interaction between continuous determinants can be estimated using logistic regression analysis.
- The methods and formulas presented in this article are intended to assist epidemiologists to calculate additive interaction between continuous determinants.
- To facilitate its use the proposed methods are included in a spreadsheet which is freely available at: www.juliuscenter.nl/additive-interaction.xls.

References

- ¹ Rothman KJ. *Epidemiology: An Introduction*. New York: Oxford University Press, 2002.
- ² Rothman KJ. *Modern Epidemiology*. Boston/Toronto: Little: Brown and Company, 1986.
- ³ Hosmer DW, Lemeshow S. Confidence interval estimation of interaction. *Epidemiology* 1992;**3**:452–56.
- ⁴ Assmann SF, Hosmer DW, Lemeshow S, Mundt KA. Confidence intervals for measures of interaction. *Epidemiology* 1996;**7**:286–90.
- ⁵ Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006;**25**:127–41.
- ⁶ Grobbee DE, Hoes AW, Verheij TJ, Schrijvers AJ, van Ameijden EJ, Numans ME. The Utrecht Health Project: optimization of routine healthcare data for research. *Eur J Epidemiol* 2005;**20**:285–87.

Appendix 1

In linear regression, the regression coefficient of the product term refers to departure from additivity, whereas in logistic regression, the regression coefficient of the product term refers to departure from multiplicativity. This is shown below. For simplicity we assume determinants *A* and *B* to be dichotomous (with levels 0 and 1).

Linear regression

When entering two determinants, *A* and *B*, and a product term in a linear regression model, the regression formula of the outcome *Y* is:

$$Y = \beta_0 + \beta_1A + \beta_2B + \beta_3AB.$$

The individual effect of A , assuming no effect of B , is:

$$Y = \beta_0 + \beta_1 - \beta_0 = \beta_1.$$

The individual effect of B , assuming no effect of A , is:

$$Y = \beta_0 + \beta_2 - \beta_0 = \beta_2.$$

The combined effect of A and B , compared with no effect of A and B , is:

$$Y = \beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 = \beta_1 + \beta_2 + \beta_3.$$

It can be seen that the combined effect of A and B can be assessed by adding the regression coefficients of A , B and the product term. There are three possibilities for the regression coefficient of the product term:

- (i) If $\beta_3 = 0$, the combined effect of A and $B = \beta_1 + \beta_2 \rightarrow$ exactly additivity \rightarrow no interaction as departure from additivity.
- (ii) If $\beta_3 < 0$, the combined effect of A and $B < \beta_1 + \beta_2 \rightarrow$ less than additivity \rightarrow ‘negative’ interaction as departure from additivity.
- (iii) if $\beta_3 > 0$, the combined effect of A and $B > \beta_1 + \beta_2 \rightarrow$ more than additivity \rightarrow ‘positive’ interaction as departure from additivity.

Logistic regression

When inserting two determinants, A and B , and a product term in a logistic regression model, the regression formula of the logit of p is:

$$\ln\left(\frac{p}{1-p}\right) = \ln(\text{odds}) = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 AB.$$

The individual effect of A , assuming no effect of B , is:

$$\begin{aligned} \ln(\text{OR}_{A+}) &= \ln\left(\frac{\text{odds}(A+,B-)}{\text{odds}(A-,B-)}\right) \\ &= \ln(\text{odds}(A+,B-)) - \ln(\text{odds}(A-,B-)) \\ &= \beta_0 + \beta_1 - \beta_0 = \beta_1 \rightarrow \text{OR}_{A+} = e^{\beta_1}. \end{aligned}$$

The individual effect of B , assuming no effect of A , is:

$$\begin{aligned} \ln(\text{OR}_{B+}) &= \ln\left(\frac{\text{odds}(A-,B+)}{\text{odds}(A-,B-)}\right) \\ &= \ln(\text{odds}(A-,B+)) - \ln(\text{odds}(A-,B-)) \\ &= \beta_0 + \beta_2 - \beta_0 = \beta_2 \rightarrow \text{OR}_{B+} = e^{\beta_2}. \end{aligned}$$

The combined effect of A and B , compared with no effect of A and B , is:

$$\begin{aligned} \ln(\text{OR}_{A+B+}) &= \ln\left(\frac{\text{odds}(A+,B+)}{\text{odds}(A-,B-)}\right) \\ &= \ln(\text{odds}(A+,B+)) - \ln(\text{odds}(A-,B-)) \\ &= \beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 = \beta_1 + \beta_2 + \beta_3 \\ &\rightarrow \text{OR}_{A+B+} = e^{\beta_1 + \beta_2 + \beta_3} = e^{\beta_1} \times e^{\beta_2} \times e^{\beta_3} \\ &= \text{OR}_A \times \text{OR}_B \times \text{OR}_{AB}. \end{aligned}$$

It can be seen that the combined effect of A and B can be assessed by multiplying OR_A , OR_B and OR_{AB} . There are three possibilities for the regression coefficient of the

product term:

- (i) If $\beta_3 = 0$, $\text{OR}_{AB} = 1$ and the combined effect of A and $B = \text{OR}_A \times \text{OR}_B \rightarrow$ exactly multiplicativity \rightarrow no interaction as departure from multiplicativity.
- (ii) If $\beta_3 < 0$, $\text{OR}_{AB} < 1$ and the combined effect of A and $B < \text{OR}_A \times \text{OR}_B \rightarrow$ less than multiplicativity \rightarrow ‘negative’ interaction as departure from multiplicativity.
- (iii) if $\beta_3 > 0$, $\text{OR}_{AB} > 1$ and the combined effect of A and $B > \text{OR}_A \times \text{OR}_B \rightarrow$ more than multiplicativity \rightarrow ‘positive’ interaction as departure from multiplicativity.

Appendix 2

Script for bootstrapping

This is a script for S-PLUS which can be used to bootstrap the RERI and its 95% confidence interval.

```
library (Design)
reri <- function(datsam)
{
  fitr <- glm(<outcome variable> ~ <variable A> *
  <variable B>, family=binomial, data=datsam)
  reri <- exp(fitr$coef[2]+fitr$coef[3]+fitr$coef[4]) -
  exp(fitr$coef[2]) - exp(fitr$coef[3]) + 1
}
summary.bootstrap(bootstrap(<dataset>, reri(<dataset>),
B = 10000, group = <dataset>$<outcome variable>), probs=c(0.025,0.5,
0.975))
```

In the general linear model (glm) command, the outcome variable and the two determinants, variable A and variable B , should be substituted. In the bootstrap command, the name of the data set should be filled in and a grouping variable should be filled in to make sure that bootstrap sampling is performed within the strata of the outcome. Furthermore, the number of samples is specified ($B = 10000$) in the script and the median and 2.5th and 97.5th percentile are asked for.

Example output

This is an example of output that S-PLUS gives, when running the script described above.

```
Forming replications 1 to 100
...
Forming replications 9901 to 10000
Call:
bootstrap(data = lrgpset, statistic = reri(lrgpset),
B = 10000, group = lrgpset$bpd.dich)
Number of Replications: 10000
Summary Statistics:
```

	Observed	Bias	Mean	SE
Reri	2.706	0.03215	2.738	0.7906
Empirical Percentiles:				
	2.5%	50%	97.5%	
Reri	1.28	2.709	4.377	
BCa Confidence Limits:				
	2.5%	50%	97.5%	
Reri	1.293	2.701	4.388	

Explanation of the output

Call

Here it says that it uses the data set 'lrgpset' and it bootstraps the statistic 'reri' with a number of samples of 10 000 000 and the outcome 'bpd.dich' as grouping variable.

Summary statistics

The observed or calculated RERI is 2.706. Of 10 000 samples the mean RERI is 2.738 and the standard error is 0.7906. The bias is the difference between the observed and the mean RERI.

Empirical percentiles

Here the median value (50th percentile) and the 2.5th and 97.5th percentiles of the RERI are given.

BCa confidence limits

These are the bias adjusted median value (50th percentile) and the bias adjusted 2.5th and 97.5th percentiles of the RERI. These 2.5th and 97.5th percentiles are the 95% CI limits. The bias adjusted median and percentiles are corrected for bias due to overfitting of the model by the bootstrap procedure.