European Journal of Epidemiology (2005) 20: 575–579
DOI 10.1007/s10654-005-7835-x

METHODS

Calculating measures of biological interaction

Tomas Andersson[1], Lars Alfredsson[1,2], Henrik Källberg[2], Slobodan Zdravkovic[2] & Anders Ahlbom[1,2]
[1]*Stockholm Centre for Public Health, Sweden;* [2]*Institute of Environmental Medicine, Stockholm, Sweden*

Accepted in revised form 24 May 2005

Abstract. An editorial in this issue explains that the degree of biological interaction between risk factors is measured as the deviation from additivity by the corresponding disease rates and not for example as deviation from multiplicativity. It is the purpose of this article to describe how a logistic regression model, or a Cox regression model, can be defined in order to produce the output that is needed for assessment of biological interaction. We will also demonstrate how common software can be programmed to deliver this output. Finally, we show how this output can be used as input in an Excel sheet that is set up to calculate the measures of biological interaction to be used for the assessment.

Key words: Biological interaction, RERI, Synergy

An editorial in this issue explains that the degree of biological interaction between risk factors is measured as the deviation from additivity by the corresponding disease rates and not for example as deviation from multiplicativity [1]. It also states that statistical interaction is quite a different thing and one that may reflect either departure from additivity or multiplicativity depending on the chosen statistical model. For an in depth discussion of the distinction between biological interaction and statistical interaction we refer for example to the recent textbook by Rothman [2].

The logistic regression model and the Cox regression model are probably the most commonly used statistical models in epidemiologic analysis to day. Because these models are exponential they are inherently multiplicative and become additive only after logarithmic transformation. Thus, absence of an interaction term in such a model implies a multiplicative relation between the disease rates and the presence of an interaction term implies departure from multiplicativity, rather than from additivity. Therefore, the interaction term, in one of these models, has no direct relevance for the issue of whether or not biological interaction is present. However, the presence of biological interaction can still be assessed from the results of a logistic regression model or a Cox regression model, but this requires that the model is defined in a special way and that the analysis is done adequately.

It is the purpose of this article to describe how a logistic regression model, or a Cox regression model, can be defined in order to produce the output that is needed for assessment of biological interaction. We will also demonstrate how common software can be programmed to deliver this output. Finally, we show how this output can be used as input in an Excel sheet that is set up to calculate the measures of biological interaction to be used for the assessment.

We will restrict our discussion to the situation with two dichotomous risk factors, in which case we have four possible combinations and, thus, four exposure categories. As for logistic regression we make the common assumption that the odds ratio can be used in lieu of the relative risk The model is set up such that it includes terms for three of the four possible combinations of exposure while the fourth category serves as reference category [3, 4]. Additional terms may be included for confounding control but have no consequences for the interaction analysis and are not discussed further in this presentation. In order to specify the model, let $i = 1$ when the first risk factor is present and 0 otherwise, and let $j = 1$ when the second is present and 0 otherwise. Furthermore, let RR_{ij} be the relative risk in exposure category i,j. Thus, RR_{11}, RR_{10}, RR_{01}, and RR_{00} are the relative risks for each of the four categories. We also define those who are unexposed to both the first and the second risk factor as reference category, i.e., $RR_{00} = 1$. There are, thus, three relative risks to be estimated and all three estimates will be required for assessment of biological interaction. Note that with this definition of the model, the relative risk for one risk factor is defined on the condition that the other risk factors are absent. For example, RR_{10} is the relative risk for the first risk factor in the absence of the second.

Based on these assumptions and definitions the three relative risk estimates can be obtained from a logistic regression model or from a Cox regression model. The corresponding covariance matrix will also be needed for calculation of confidence intervals. In order to obtain the adequate estimates, the model is set up with indicator variables for each of the four different combinations of exposure. A convenient way to do this is to define three indicator variables ind11, ind10, and ind01 as in Table 1.

Appendix 1 provides a SAS program that defines the model and delivers estimates of the required parameters together with the covariance matrix. The appendix first demonstrates how to do this for logistic regression. There is then an instruction for how to adjust the program for use of the Cox regression model. Appendix 2 explain how to achieve the same things in STATA and Appendix 3, how to do this for logistic regression in SPSS.

Rothman presents three measures of biological interaction: RERI, the relative excess risk due to interaction; AP, the attributable proportion due to interaction; and S, the synergy index [5]. These measures are defined as follows:

$$\mathrm{RERI} = \mathrm{RR}_{11} - \mathrm{RR}_{10} - \mathrm{RR}_{01} + 1,$$

$$\mathrm{AP} = \mathrm{RERI}/\mathrm{RR}_{11},$$

$$S = [\mathrm{RR}_{11} - 1]/[(\mathrm{RR}_{10} - 1) + (\mathrm{RR}_{01} - 1)]$$

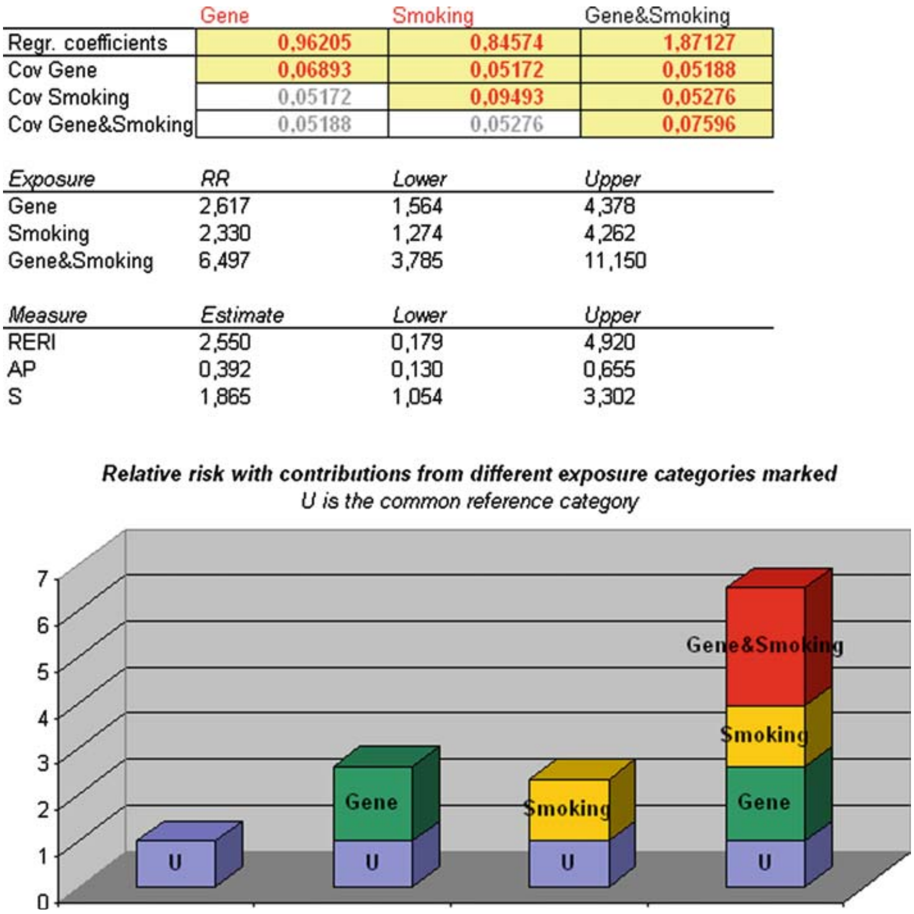
Table 1. Definition of dummy variables for different exposure combinations

Exposure levels	ind01	ind10	ind11
$i = 0, j = 0$	0	0	0
$i = 0, j = 1$	1	0	0
$i = 1, j = 0$	0	1	0
$i = 1, j = 1$	0	0	1

If there is no biological interaction, RERI and AP are equal to 0 and S is equal to 1. The estimation of these measures is straightforward assuming that estimates of RR_{11}, RR_{10}, and RR_{01} are available. For the calculation of confidence intervals for the three measures of biological interaction there are several possibilities [6]. We use what Hosmer and Lemeshow refer to as the delta method [3], which is a straightforward Taylor expansion of the variances and covariances [4].

Biological interaction in Excel

We provide here access to an Excel sheet which can be used to calculate the three measures of interaction



	Gene	Smoking	Gene&Smoking
Regr. coefficients	0,96205	0,84574	1,87127
Cov Gene	0,06893	0,05172	0,05188
Cov Smoking	0,05172	0,09493	0,05276
Cov Gene&Smoking	0,05188	0,05276	0,07596

Exposure	RR	Lower	Upper
Gene	2,617	1,564	4,378
Smoking	2,330	1,274	4,262
Gene&Smoking	6,497	3,785	11,150

Measure	Estimate	Lower	Upper
RERI	2,550	0,179	4,920
AP	0,392	0,130	0,655
S	1,865	1,054	3,302

Relative risk with contributions from different exposure categories marked
U is the common reference category
0
1
2
3
4
5
6
7
U
Gene
U
Smoking
U
Gene&Smoking
Smoking
Gene
U

Figure 1. Excel sheet for calculating measures of biological interaction.

and their confidence intervals based on the results from a logistic regression or Cox regression model. This program can be found at: www.epinet.se.

This Excel sheet requires as input the regression coefficients for each of the three exposure categories. Most statistical software provides both the regression coefficients and the odds ratios (OR) or hazard ratios (HR) and both could be used for these calculations, but this particular Excel sheet expects the regression coefficients as input. The Excel sheet also requires the covariance matrix in order to be able to carry out the confidence interval calculations. Most software will provide the covariance matrix for the regression coefficients, but normally only if requested. The programs in the Appendix 1–3 include instructions for all the required estimates to be delivered as output. Because there are three estimated regression coefficients there are six pair-wise combinations each having a variance or a covariance. So the Excel sheet expects a total of nine numbers as input, three regression coefficients and six covariances.

Figure 1 shows an example of how to use the Excel sheet. The example investigates the possible biological interaction between one gene and smoking in relation to rheumatoid arthritis [7]. The nine numbers in the first panel's upper right hand corner in bold print (red on the actual sheet) are filled in by the user and the rest is calculated by Excel that also provides the figure. The second panel in the table displays the three estimated relative risks with 95% confidence intervals. The last panel gives the three measures of biological interaction, also with 95% confidence intervals. All the three measures of biological interaction indicate that there is indeed a biological interaction between this particular gene and smoking. It is worth noting that despite this, the regression coefficients are in essence additive, as seen in the first row of the first panel and the relative risks are almost multiplicative as seen in the second panel of the table. In the figure, U represents the risk in the reference category and is defined as 1. The three other bars in the figure illustrate the relative excess risk due to the gene, the relative excess risk due to smoking, and the relative excess risk due to the biological interaction between the gene and smoking.

The measures of biological interaction can of course also be calculated directly in SAS, STATA, or SPSS without using this Excel sheet, and indeed the reference by Lundberg includes a program describing how SAS does this for logistic regression [3]. However, some may find the approach presented here more accessible.

Appendix 1.

Logistic regression in SAS

```
/* This is an example of a SAS program
using proc logistic to generate the input
needed in the excel sheet. The aim of
this model is to calculate the biological
interaction between smoking and a gene in
relation to Rheumatoid Arthritis with
control for sex and age in categories. */

  /* We can't use the original indicator
variable coding because we need four
disjoint categories (non-smoker and non-
gene, non-smoker and gene, smoker and
non-gene, smoker and gene) and thus three
indicator variables. */

DATA d1;
SET d1;
/* Make sure that no missing gets
coded as 00 */
IF smoker^ =. AND gene^ =. THEN DO;
/* non-smoker and gene */
IF smoker = 0 AND gene = 1
  THEN ind01 = 1;
ELSE ind01 = 0;
/* smoker and non-gene */
IF smoker = 1 AND gene = 0
  THEN ind10 = 1;
ELSE ind10 = 0;
/* smoker and gene */
IF smoker = 1 AND gene = 1
  THEN ind11 = 1;
ELSE ind11 = 0;
END;

/* Save the parameter estimates and the
covariance matrix from proc logistic
into d2 */
PROC LOGISTIC DATA = d1 DESCENDING
  OUTEST = d2 COVOUT;
MODEL case = ind01 ind10 ind11 sex
  age_dum1-age_dum9;


/* Remove not needed information
from d2 */
DATA d2;
SET d2;
IF _name_ in(‘ind01’,‘ind10’,‘ind11’)
  OR _type_=‘PARMS’;
KEEP _name_ ind01 ind10 ind11;


/* And finally a print of the numbers
needed for the excel sheet */
PROC PRINT DATA = d2;
```

Cox regression in SAS

```
/* Use the same code as above and replace
the proc logistic statement with this */
```

```
PROC PHREG DATA = d1 OUTEST = d2 COVOUT;
MODEL time*case(0) = ind01 ind10 ind11
  sex age_dum1-age_dum9;
```

Appendix 2.

Logistic regression in STATA

```
/* This is an example of a STATA program
using proc logistic to generate the input
needed in the excel sheet. The aim of
this model is to calculate the biological
interaction between smoking and a gene in
relation to Rheumatoid Arthritis with
control for sex and age in categories. */

  /* We can't use the original indicator
variable coding because we need four
disjoint categories (non-smoker and non-
gene, non-smoker and gene, smoker and
non-gene, smoker and gene) and thus three
indicator variables. */



gen ind11 = 1      if smoker == 1 & gene == 1
gen ind10 = 1      if smoker == 1 & gene == 0
gen ind01 = 1      if smoker == 0 & gene == 1
replace ind11 = 0   if ind11==.
replace ind10 = 0   if ind10==.
replace ind01 = 0   if ind01==.
/* Make sure no that no missing gets
coded as 00 */
replace ind11=.   if smoker==. gene==.
replace ind10=.   if smoker==. gene==.
replace ind01=.   if smoker==. gene==.

/* logistic regression*/
/* dependent variable coded as 1 = case
0 = control */
logistic case ind01 ind10 ind11 sex
  age_dum1 age_dum2 age_dum3 age_dum4
  age_dum5 age_dum6 age_dum7 age_dum8
  age_dum9, coef

/* And finally the covariance matrix and
       numbers needed for the excel sheet */
Vce
```

Cox regression in STATA

```
/* Use the same code as above and replace
the logistic statement with this */
stset time, failure(case) stcox ind01
    ind10 ind11 sex age_dum1 age_dum2
    age_dum3 age_dum4 age_dum5 age_dum6
    age_dum7 age_dum8 age_dum9, nohr
```

Appendix 3.

Logistic regression in SPSS

```
* This is an example of a SPSS program
* using NOMREG to generate the input
* needed in the excel sheet. The aim of
* this model is to calculate the
* biological interaction between smoking
* and a gene in relation to Rheumatoid
* Arthritis with control for sex and age
* in categories.

* We can't use the original indicator
* variable coding because we need four
* disjoint categories (non-smoker and
* non-gene, non-smoker gene, smoker and
* non-gene, smoker gene) and thus three
* indicator variables.


* Make sure that no missing gets
* coded as 00
* Syntax
IF (smoker = 1 and gene = 1) ind11 = 1 .
EXECUTE.


IF ((smoker = 1 and gene = 0) or
  (smoker = 0 and gene = 1) or
  (smoker = 0 and gene = 0)) ind11 = 0.
EXECUTE.
IF (smoker = 1 and gene = 0) ind10 = 1.
EXECUTE.
IF ((smoker = 1 and gene = 1) or
  (smoker = 0 and gene = 1) or
  (smoker = 0 and gene = 0)) ind10 = 0.
EXECUTE.
IF (smoker = 0 and gene = 1) ind01 = 1.
EXECUTE.
IF ((smoker = 1 and gene = 1) or
  (smoker = 1 and gene = 0) or
  (smoker = 0 and gene = 0))ind01 = 0.
EXECUTE.


* logistic regression (ANALYZE/REGRESSION
* /MULTINOMIAL (ask for covariances in
* statistics)
* Dependent variable coded as 1 = case
* 0 = control
* Syntax
NOMREG
case (BASE = FIRST ORDER = ASCENDING)
WITH ind01 ind10 ind11 sex age_dum1
  age_dum2 age_dum3 age_dum4 age_dum5
  age_dum6 age_dum7 age_dum8 age_dum9
/INTERCEPT = INCLUDE
/PRINT = COVB PARAMETER.
```

References

1. Ahlbom A, Alfredsson L. Interaction: Word with two meanings creates confusion. Eur J Epidemiol, in press.
2. Rothman KJ. Epidemiology. An Introduction. New York: Oxford University Press, 2002.
3. Lundberg M, Fredlund P, Hallqvist J, Diderichsen F. A SAS program calculating three measures of interaction with confidence intervals. Epidemiology 1996; 7: 655–656.
4. Hosmer DW, Lemeshow S. Confidence interval estimation of interaction. Epidemiology 1992; 3: 452–456.
5. Hallqvist J, Ahlbom A, Diderichsen F, Reuterwall C. How to evaluate interaction between causes: A review of practices in cardiovascular epidemiology. J Intern Med 1996; 239: 377–382.
6. Assmann SF, Hosmer DW, Lemeshow S, Mundt KA. Confidence intervals for measures of interaction. Epidemiology 1996; 7: 286–290.
7. Padyukov L, Silva C, Stolt P, Alfredsson L, Klareskog L. A gene–environment interaction between smoking and shared epitope genes in HLA-DR provides a high risk of seropositive rheumatoid arthritis. Arthritis Rheum 2004; 50: 3085–3092.

Address for correspondence: Tomas Andersson, Stockholm Centre for Public Health, Norrbacka Building, Karolinska University Hospital, 171 76 Stockholm, Sweden
E-mail: tomas.andersson@imm.ki.se