Measuring and estimating interaction between exposures on dichotomous outcome of a population

Xiaoqin Wang\textsuperscript{1}, Weimin Ye\textsuperscript{2} and Li Yin\textsuperscript{2,*}

\textsuperscript{1}Department of Electronics, Mathematics and Natural Sciences, University of Gävle, 801 76, Gävle, Sweden
\textsuperscript{2}Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Box 281, SE 171 77, Stockholm, Sweden
\textsuperscript{*}Corresponding author. E-mail: Li.Yin@ki.se, Phone: +46-8-52486187, Fax: +46-8- 311101

Summary: In observational studies for the interaction between exposures on dichotomous outcome of a population, one usually uses one parameter of a regression model to describe the interaction, leading to one measure of the interaction. In this article, we use the conditional risk of outcome given exposures and covariates to describe the interaction and obtain five different measures for the interaction in observational studies, i.e. difference between the marginal risk differences, ratio of the marginal risk ratios, ratio of the marginal odds ratios, ratio of the conditional risk ratios, and ratio of the conditional odds ratios. By using only one regression model for the conditional risk of outcome given exposures and covariates, we obtain the maximum-likelihood estimates of all these measures. By generating approximate distributions of the maximum-likelihood estimates of these measures, we obtain interval estimates of these measures. The method is presented by studying the interaction between a therapy and the environment on eradication of \textit{Helicobacter pylori} among Vietnamese children.

\textbf{Key words:} Interaction; Interaction measure; Point estimate; Interval estimate; Regression model

Subject classification codes: 62F25, 62E17, 62F10

1. Introduction

The interaction between exposures on an outcome refers to the situation where the effect of one exposure on the outcome depends on the assignment of another exposure \cite{21, 22, 24}. Suppose a hypothetical randomized trial in which one could assign two exposures $z_1$ and $z_2$ and wished to investigate the effect of two exposures $z_1$ and $z_2$ on an outcome $y$ of certain population, where the variables $z_1, z_2$ and $y$ are all dichotomous, i.e. $z_1 = 0, 1$, $z_2 = 0, 1$ and $y = 0, 1$. In the randomized trial, covariates are essentially unassociated with exposures $(z_1, z_2)$ and thus are not confounders. In this case, one may use the marginal risk $\text{pr}(y = 1 \mid z_1, z_2)$ to measure the effect of $z_1$ and then the interaction between $z_1$ and $z_2$. Here the marginal risk $\text{pr}(y = 1 \mid z_1, z_2)$ is marginal with respect to covariates and thus conditional on exposures $(z_1, z_2)$ only.

One may measure the effect of $z_1$ by the marginal risk difference
\[
\text{MRD}(z_2) = \text{pr}(y = 1 \mid z_1 = 1, z_2) - \text{pr}(y = 1 \mid z_1 = 0, z_2)
\]
and the interaction by difference between the marginal risk differences
\[
\text{DMRD} = \text{MRD}(z_2=1) - \text{MRD}(z_2=0).
\]

One may also measure the effect of $z_1$ by the marginal risk ratio
\[
\text{MRR}(z_2) = \frac{\text{pr}(y = 1 \mid z_1 = 1, z_2)}{\text{pr}(y = 1 \mid z_1 = 0, z_2)}
\]
and the interaction by ratio of the marginal risk ratios
RMRR = \frac{MRR(z_2=1)}{MRR(z_2=0)}.

One may also measure the effect of \( z_1 \) by the marginal odds ratio

\[
\text{MOR}(z_2) = \frac{\text{pr}(y = 1 | z_1 = 1, z_2)}{\text{pr}(y = 1 | z_1 = 0, z_2)} \cdot \frac{1 - \text{pr}(y = 1 | z_1 = 1, z_2)}{1 - \text{pr}(y = 1 | z_1 = 0, z_2)}
\]

and the interaction by ratio of the conditional odds ratios

\[
\text{RMOR} = \frac{\text{MOR}(z_2=1)}{\text{MOR}(z_2=0)}.
\]

Oftentimes, DMRD is called the biological interaction because its expression in terms of probability differences has a close relationship with some well-known classification of biological mechanisms [7, 21]. However, the other two measures RMRR and RMOR of the interaction are also commonly used, have causal interpretation in the framework of Rubin causal model [19, 20, 23], and reflect different aspects of the underlying the interaction [24].

Now suppose an observational study in which one investigates the effect of the exposure \((z_1, z_2)\) on the outcome \(y\) of the population above. In the observational study, covariates may be associated with both the outcome \(y\) and the exposure \((z_1, z_2)\) and therefore are confounders. For illustrative purpose, we consider the case of one cofounder \(x\). To adjust for \(x\) in estimation of the effect of \((z_1, z_2)\), one needs to model the conditional risk \(\text{pr}(y = 1 | z_1, z_2, x)\) given not only \((z_1, z_2)\) but also \(x\). When sample size is small, one often obtains a simple model for \(\text{pr}(y = 1 | z_1, z_2, x)\), for instance,

\[
g\{\text{pr}(y = 1 | z_1, z_2, x)\} = \alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 (z_1 * z_2) + \theta_1 x
\]

where \(g(.)\) is a link function. The model contains one term for the product \(z_1 * z_2\), but no terms for the products between \(x\) and \((z_1, z_2)\). The parameter \(\beta_3\) is called the statistical interaction which describes deviation of this model from the main effect model containing no term for any product between variables [7, 21, 22].

If there are no unmeasured confounders than the confounder \(x\), then \(\beta_3\) or a function of \(\beta_3\) has causal interpretation and measures the interaction on any stratum \(x = 0, 1\) and thus on the population. With the identity link function \(g(p) = p\), the parameter \(\beta_3\) is difference between the conditional risk differences (DCRD), which is also equal to difference between the marginal risk differences (DMRD) because risk differences are collapsible. With the log link function \(g(p) = \log(p)\), the exponential \(\exp(\beta_3)\) is ratio of the conditional risk ratios (RCRR), which is also equal to ratio of the marginal risk ratios (RMRR). With the logistic link function \(g(p) = \logit(p)\), the exponential \(\exp(\beta_3)\) is ratio of the conditional odds ratios (RCOR), but which is not equal to ratio of the marginal odds ratios (RMOR) due to non-collapsibility of odds ratios. See, for instance, [5, 10, 12, 14] for collapsibility of measures of the exposure effect. In the method above, one first has a model and then uses a parameter of the model to measure the interaction. As a consequence, one uses one model to estimate only one measure of the interaction.

When sample size is large, the model inevitably becomes complex, for instance,

\[
g\{\text{pr}(y = 1 | z_1, z_2, x)\} = \alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 (z_1 * z_2) + \theta_1 x + \theta_2 (z_1 * x) + \theta_3 (z_2 * x)
\]

which contains, in addition to one term for the product \(z_1 * z_2\), two terms for the products \(z_1 * x\) and \(z_2 * x\). The parameter \(\beta_3\) or a function of \(\beta_3\) still measures the interaction. With a linear function, \(\beta_3\) is DCRD which is equal to DMRD. With a log link function, \(\exp(\beta_3)\) is RCRR but no longer equal to RMRR. With a logistic link function, \(\exp(\beta_3)\) is RCOR but still not equal to RMOR. In practical researches, one usually omits the last two terms in the model despite \(\theta_2 \neq 0\) and \(\theta_3 \neq 0\), leading to a
mis-specified model for \( \text{pr}(y = 1|z_1, z_2, x) \). Then the parameter \( \beta_3 \) does not measure the interaction. Furthermore, the model may even contain a triple-product term \( \theta_4(z_1 \ast z_2 \ast x) \). In this case, the parameter \( \beta_3 \) measures the interaction on stratum \( x = 0 \) but not on the population. If one omits this triple-product term despite \( \theta_4 \neq 0 \), one obtains a mis-specified model for \( \text{pr}(y = 1|z_1, z_2, x) \), but then, \( \beta_3 \) does not measure the interaction on stratum \( x = 0 \) nor on the population. With these mis-specified models, it would not be possible to have consistent estimate of any measure of the interaction.

In this article, we measure and estimate the interaction between exposures in two separate steps. In the measuring step, we express all the five measures of the interaction, DMRD, RMRR, RMOR, RCRR and RCOR, in terms of the conditional risk \( \text{pr}(y = 1|z_1, z_2, x) \), rather than model parameters, where \( x \) is a vector of covariates. In the estimating step, we estimate DMRD, RMRR, RMOR, RCRR and RCOR through \( \text{pr}(y = 1|z_1, z_2, x) \), which can be described by any model. As a result, we can use one model in one study to estimate all these measures.

Our method of measuring and estimating the interaction is extension of a well-known method of measuring and estimating the exposure effect. In the latter method, one expresses a measure of the exposure effect in terms of \( \text{pr}(y = 1|z_1, z_2, x) \) and then estimates the measure through \( \text{pr}(y = 1|z_1, z_2, x) \) [1, 6, 8, 13]. The difficulty with this method is interval estimation for the measure of the exposure effect. The common method for the interval estimation is the parametric or non-parametric bootstrap method, but it is highly difficult to correct the finite-sample bias arising from the bootstrap sampling with dichotomous outcome [3, 4, 9]. To avoid complexity of the bootstrap method for interval estimation of a measure of the interaction, we use distribution of the ML estimates of the model parameters to obtain distribution of the ML estimate of the measure and then the corresponding confidence interval.

We are going to present the method through an observational study embedded in a randomized trial, in which we investigated the interaction between a therapy and the environment on eradication of \textit{Helicobacter pylori} among Vietnamese children.

2. The interaction between a therapy and the environment on eradication of \textit{Helicobacter pylori} among Vietnamese children

2.1 Medical background and the data

In a randomized trial [18], researchers studied two triple therapies – (lansoprazole, amoxicillin, metronidazole) and (lansoprazole, amoxicillin, clarithromycin) – for their abilities to eradicate \textit{Helicobacter pylori} among Vietnamese children. From several children hospitals in Hanoi, restricting body weight to a range between 13 kg and 45 kg, a sample was collected between May 2005 and January 2006. In an observational substudy embedded in this randomized trial, they focused on one treatment arm of the triple therapy (lansoprazole, amoxicillin, metronidazole). They analyzed the effect of high versus low doses of the therapy and the effect of the environment encoded by geographic areas in which these children lived. In this article, we study how the environment interacted with the therapy on eradication of \textit{Helicobacter pylori} among Vietnamese children. The treatment arm comprised 109 patients.

The therapy eradicated \textit{Helicobacter pylori} through systemic circulation, so the researchers assigned the therapy to the children according to their body weights. According to the pediatric procedure, children with 13 kg < body weight < 23 kg received the therapy once daily and those with 23 kg ≤ weight < 45 kg twice daily. Of medical relevance was dose in unit body weight. Following
the earlier report on the same data [18], we categorized the children into receiving low versus high doses of the therapy by taking the middle point of each weight category as the threshold.

Hence children with 13 kg < body weight ≤ 18 kg in the first weight category and 23 kg ≤ body weight < 34 kg in the second weight category were considered as receiving the high dose \( z_1 = 1 \) of the therapy whereas those with 18 kg < body weight < 23 kg in the first weight category and 34 kg ≤ body weight < 45 kg in the second weight category as receiving the low dose \( z_1 = 0 \). Let \( z_2 = 1 \) indicate the rural environment and \( z_2 = 0 \) the urban environment.

Then the exposures in this study were \( z = (z_1, z_2) = (\text{therapy, environment}) \). The outcome was successful \( (y = 1) \) versus unsuccessful \( (y = 0) \) eradication of *Helicobacter pylori*. In this article, we estimate the interaction between exposures \( z_1 \) and \( z_2 \) on the outcome \( y = 1 \) among Vietnamese children.

In the context of this study, body weight was synonymous to exposure \( z_1 \) and thus was not an independent covariate. The covariate age \( (x_1) \) was associated with body weight and thus with the assignment of exposure \( z_1 \). Furthermore, age was known to have an effect on the outcome \( y \) and thus was a confounding covariate. In addition to age, other possible confounding covariates were recorded, which were gender \( (x_2) \) and antibiotic resistance to metronidazole \( (x_3) \). Age was categorized into younger \( (x_1 = 1) \) versus older \( (x_1 = 0) \) than 9 years. Let \( x_2 = 1 \) indicate female and \( x_2 = 0 \) male of the gender. Antibiotic resistance to metronidazole was categorized into non-resistant \( (x_3 = 1) \) versus resistant \( (x_3 = 0) \). Let \( x = (x_1, x_2, x_3) \) be the set of all the recorded covariates. The data of this study is given in Table 1.

### 2.2 Supposition for interaction of causal interpretation

In this article, we aim at the interaction of causal interpretation, that is, we compare the potential outcomes \( y(z) \) of the patients in the population under different exposures \( z = (z_1, z_2) = (0, 0), (0, 1), (1, 0), (1, 1) \). Because it is only possible to observe potential outcome of a patient under one value of the exposure \( z \), the following supposition is needed to allow for estimation of the interaction [19, 20, 23].

**Supposition** The assignment of exposures \( z \) is strongly ignorable given the covariates \( x \) and the probability \( \Pr(z|x) \) of \( z \) given \( x \) is larger than zero. Therefore the risk \( \Pr(y|z,x) \) of the potential outcome \( y(z) \) in stratum \( x \) under exposure \( z \) is equal to the risk \( \Pr(y|z,x) \) of the observable outcome \( y \) in stratum \( (x,z) \), that is, \( \Pr(y(z)|x) = \Pr(y|z,x) \).

The supposition is also called the assumption of no unmeasured confounding covariates given the covariates \( x \) [11, 21]; see [24] for examples of the supposition and their causal directed acyclic graphs.

### 3. Measuring interaction on population

#### 3.1 Conditional measures of interaction on population

The conditional odds \( \text{CO}(z; x) \) of the potential outcome \( y(z) = 1 \) in stratum \( x = (x_1, x_2, x_3) \) under exposure \( z = (z_1, z_2) \) is

\[
\text{CO}(z; x) = \frac{\Pr(y(z) = 1|x)}{\Pr(y(z) = 0|x)},
\]

where \( \Pr(y(z) = 0|x) = 1 - \Pr(y(z) = 1|x) \).
The conditional odds ratio is

\[ \text{COR}(z_2; x) = \frac{\text{CO}(z_1 = 1, z_2; x)}{\text{CO}(z_1 = 0, z_2; x)}, \tag{2} \]

which measures the effect of the therapy \( z_1 \) on stratum \( x \) in which each patient were assigned environment \( z_2 \). Then the ratio of the conditional odds ratios is

\[ \text{RCOR}(x) = \frac{\text{COR}(z_2 = 1; x)}{\text{COR}(z_2 = 0; x)}, \tag{3} \]

which measures the interaction between the exposures \( z_1 \) and \( z_2 \) on stratum \( x \). Taking the average of RCOR(\( x \)) over \( x \) of the population, we obtain the ratio of the conditional odds ratio on the population

\[ \text{RCOR} = \sum_x \text{pr}(x) \text{RCOR}(x), \tag{4} \]

which measures the interaction on the population on the conditional odds ratio scale. In particular, if RCOR(\( x \)) takes the same value for all \( x \) values, then RCOR(\( x \)) = RCOR. Under Supposition in section 2.2, we have \( \text{pr}\{y(z)|x\} = \text{pr}(y|z,x) \). Inserting this equality into (1) and then using (2)-(4), we obtain

\[ \text{RCOR} = \sum_x \text{pr}(x) \text{RCOR}(x), \tag{5} \]

The conditional risk ratio is

\[ \text{CRR}(z_2; x) = \frac{\text{pr}\{y(z_1 = 1, z_2 = 1)|x\}}{\text{pr}\{y(z_1 = 0, z_2 = 1)|x\}}, \tag{6} \]

which measures the effect of \( z_1 \) on stratum \( x \) in which each patient were assigned \( z_2 \). The ratio of the conditional risk ratios is

\[ \text{RCRR}(x) = \frac{\text{CRR}(z_2 = 1; x)}{\text{CRR}(z_2 = 0; x)}, \tag{7} \]

which measures the interaction on stratum \( x \). Taking the average of RCRR(\( x \)) over \( x \) of the population, we obtain the ratio of the conditional risk ratios on the population

\[ \text{RCRR} = \sum_x \text{pr}(x) \text{RCRR}(x), \tag{8} \]

which measures the interaction on the population on the conditional risk ratio scale. In particular, if RCRR(\( x \)) takes the same value for all \( x \) values, then RCRR(\( x \)) = RCRR. Inserting the equality \( \text{pr}\{y(z)|x\} = \text{pr}(y|z,x) \) into (6) and then using (6)-(8), we obtain

\[ \text{RCRR} = \sum_x \text{pr}(x) \text{RCRR}(x), \tag{9} \]

We may also use the conditional risk difference

\[ \text{CRD}(z_2; x) = \text{pr}\{y(z_1 = 1, z_2 = 1)|x\} - \text{pr}\{y(z_1 = 0, z_2 = 1)|x\} \tag{10} \]
to measure the effect of \( z_1 \) on stratum \( x \) in which each patient were assigned \( z_2 \), and the difference between the conditional risk differences

\[
DCRD(x) = CRD(z_2 = 1; x) - CRD(z_2 = 0; x)
\]  

(11)
to measure the interaction on stratum \( x \), and the average of \( DCRD(x) \) over \( x \) of the population

\[
DCRD = \sum_x pr(x)DCRD(x)
\]  

(12)
to measure the interaction on the population on the conditional risk difference scale. In particular, if \( DCRD(x) \) takes the same value for all \( x \) values, then \( DCRD = DCRD \). Inserting the equality \( pr(y(z)|x) = pr(y | z, x) \) into (10) and then using (10)-(12), we obtain

\[
DCRD = \sum_x pr(x) \{ [pr(y = 1 | z_1 = 1, z_2 = 1, x) - pr(y = 1 | z_1 = 0, z_2 = 1, x)] \\
- [pr(y = 1 | z_1 = 1, z_2 = 0, x) - pr(y = 1 | z_1 = 0, z_2 = 0, x)] \}
\]  

(13)

3.2 Marginal measures of interaction on population

The population risk is the risk \( pr\{y(z) = 1\} \) of the potential outcome \( y(z) = 1 \) of the population under exposure \( z = (z_1, z_2) \). We rewrite the population risk \( pr\{y(z) = 1\} \) by

\[
pr\{y(z) = 1\} = \sum_x pr(y(z) = 1|x) pr(x).
\]

Under Supposition in section 2.2, we have \( pr\{y(z) = 1|x\} = pr(y = 1 | z, x) \). Therefore we obtain

\[
pr\{y(z) = 1\} = \sum_x pr(y = 1 | z, x) pr(x).
\]

Let \( PR(y = 1 | z) = \sum_x pr(y = 1 | z, x) pr(x) \), which is called the population-adjusted risk [6, 8, 13]. Then under Supposition we have that the population-adjusted risk is same as the population risk, namely

\[
PR(y = 1 | z) = pr\{y(z) = 1\} = \sum_x pr(y = 1 | z, x) pr(x).
\]  

(14)

The marginal odds \( MO(z) \) of the potential outcome \( y(z) = 1 \) of the population under exposure \( z \) is

\[
MO(z) = \frac{pr\{y(z) = 1\}}{pr\{y(z) = 0\}},
\]  

(15)

where \( pr\{y(z) = 0\} = 1 - pr\{y(z) = 1\} \). The marginal odds ratio is

\[
MOR(z_2) = \frac{MO(z_1 = 1, z_2)}{MO(z_1 = 0, z_2)},
\]  

(16)

which measures the effect of the therapy \( z_1 \) on the population in which all patients were assigned environment \( z_2 \). The ratio of marginal odds ratios is

\[
RMOR = \frac{MOR(z_2 = 1)}{MOR(z_2 = 0)}.
\]  

(17)
which measures the interaction between the exposures $z_1$ and $z_2$ on the population on the marginal odds ratio scale. Using formulas (14)-(17), we obtain

$$RMOR = \frac{\text{PR}(y = 1 \mid z_1 = 1, z_2 = 1)\text{PR}(y = 0 \mid z_1 = 0, z_2 = 1)}{\text{PR}(y = 0 \mid z_1 = 1, z_2 = 1)\text{PR}(y = 1 \mid z_1 = 0, z_2 = 1)},$$ \hfill (18)

The marginal risk ratio is

$$MRR(z_2) = \frac{\text{pr}\{y(z_1 = 1, z_2) = 1\}}{\text{pr}\{y(z_1 = 0, z_2) = 1\}}, \hfill (19)$$

which measures the effect of $z_1$ on the population in which each patient were assigned $z_2$. The ratio of marginal risk ratios is

$$RMRR = \frac{MRR(z_2 = 1)}{MRR(z_2 = 0)}, \hfill (20)$$

which measures the interaction on the population on the marginal risk ratio scale. Using (14), (19) and (20), we obtain

$$RMRR = \frac{\text{PR}(y = 1 \mid z_1 = 1, z_2 = 1)}{\text{PR}(y = 1 \mid z_1 = 0, z_2 = 1)}.$$ \hfill (21)

The marginal risk difference is

$$MRD(z_2) = \text{pr}\{y(z_1 = 1, z_2) = 1\} - \text{pr}\{y(z_1 = 0, z_2) = 1\}, \hfill (22)$$

which measures the effect of $z_1$ on the population in which each patient were assigned $z_2$. The difference of the marginal risk differences is

$$DMRD = MRD(z_2 = 1) - MRD(z_2 = 0), \hfill (23)$$

which measures the interaction on the population on the marginal risk difference scale. Using formulas (14), (22) and (23), we obtain

$$DMRD = \text{PR}(y = 1 \mid z_1 = 1, z_2 = 1) - \text{PR}(y = 1 \mid z_1 = 0, z_2 = 1) - \text{PR}(y = 1 \mid z_1 = 1, z_2 = 0) + \text{PR}(y = 1 \mid z_1 = 0, z_2 = 0). \hfill (24)$$

In randomized trial, the covariates $x$ are essentially unassociated with the exposure $z = (z_1, z_2)$, i.e. $\text{pr}(x \mid z) = \text{pr}(x)$. As a result, formula (14) becomes

$$\text{pr}\{y(z) = 1\} = \text{PR}(y = 1 \mid z) = \text{pr}(y = 1 \mid z),$$

where $\text{pr}(y = 1 \mid z)$ is the marginal risk of the outcome $y = 1$ in stratum $z$. In this case, the marginal measures of the interaction can be obtained from the marginal risk $\text{pr}(y = 1 \mid z)$ instead of $\text{pr}(y = 1 \mid z, x)$. Therefore the marginal measures we have obtained in this subsection are the same as those we described in the introduction for randomized trial. However, in many situations like the example of this article, randomized trial is not possible, and one can only rely upon observational studies and therefore needs to use the population-adjusted risk $\text{PR}(y = 1 \mid z)$ to obtain marginal measures of the interaction on population.

3.3 Properties of measures of interaction on population
From (4), we see that RCOR compares the potential outcomes \( y(z) \) of the patients in the population under different exposures \( z = (z_1, z_2) = (0, 0), (0, 1), (1, 0), (1, 1) \) and thus has causal interpretation in the framework of Rubin Causal Model (see, for instance, [19, 20, 23]). Similarly, from (8), (17), (20) and (23), we see that RCRR, RMOR, RMRR and DMRD all have causal interpretation.

These measures are symmetric: we obtain the same formulas for them if we switch the order of exposures \( z_1 \) and \( z_2 \). We have RCOR \( \neq \) RMOR and RCRR \( \neq \) RMRR because the odds ratio and the risk ratio are not collapsible [5, 10, 12, 14]. We always have DCRD = DMRD because the risk difference is collapsible. Thus we have five different measures of the interaction.

These measures reflect different aspects of the same underlying interaction: the null hypotheses, i.e. RCOR = 1, RCRR = 1, RMOR = 1, RMRR = 1 and DMRD = 0, do not imply one another. The conditional measures RCOR and RCRR describe the interaction on population in terms of the conditional risk \( \text{pr}(y = 1 \mid z, x) \), emphasizing the interaction on individual covariate-specific strata of the population. The marginal measures RMOR, RMRR and DMRD express interaction on population in terms of the population risk \( \text{pr}(y = 1) \), emphasizing the interaction on the population as a whole.

According to formulas (5) and (9), the conditional measures RCOR and RCRR can be expressed in terms of \( \text{pr}(y = 1 \mid z, x) \) and \( \text{pr}(x) \). According to (18), (21) and (24), the marginal measures RMOR, RMRR and DMRD can be expressed in terms of \( \text{PR}(y = 1 \mid z) \) which in turn can be expressed in terms of \( \text{pr}(y = 1 \mid z, x) \) and \( \text{pr}(x) \) according to (14). Therefore all the five measures can be expressed in terms of \( \text{pr}(y = 1 \mid z, x) \) and \( \text{pr}(x) \), implying that we can estimate these measures by estimating \( \text{pr}(y = 1 \mid z, x) \) and \( \text{pr}(x) \) through observed data. In other words, we can use any model to estimate these measures if the model is correctly specified for \( \text{pr}(y = 1 \mid z, x) \).

4. Estimating interaction on population

4.1 Regression model

By fitting the data introduced in Section 2 to a logistic model, we obtain the following model for the risk \( \text{pr}(y = 1 \mid z, x) \) of \( y = 1 \) in stratum \( (z, x) \)

\[
\log \left( \frac{\text{pr}(y = 1 \mid z, x)}{1 - \text{pr}(y = 1 \mid z, x)} \right) = \alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 (z_1 \ast z_2) + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 (z_1 \ast x_2). \tag{25}
\]

In this model, we include, in addition to one term for the product \( z_1 \ast z_2 \), another term for the product \( z_1 \ast x_2 \), because of the small p-value, 0.10, for the likelihood ratio-based significance test of \( \theta_4 = 0 \). Let \( \pi = (\alpha, \beta_1, \beta_2, \beta_3, \theta_1, \theta_2, \theta_3, \theta_4) \) be the set of all model parameters. The ML estimate \( \hat{\pi} = (\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4) \) and its approximate covariance matrix \( \hat{\Sigma} \) (i.e. the inverse of the observed information) are given in Table 2.

Combining model (25) with formula (9), we have that \( \exp(\beta_3) \) is equal to RCOR. In addition to RCOR, we are also interested in other aspects of the interaction on the population described by RCRR, RMOR, RMRR and DMRD, which are not functions of \( \beta_3 \). In particular, since the outcome \( y \) is common as seen in Table 1, the conditional odds ratios are poor approximations of the conditional risk ratios, so we cannot convert conditional odds ratios into other measures of interaction. If we use a log-linear model

\[
\log \{\text{pr}(y = 1 \mid z, x)\} = \]
\[
\alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 (z_1 \ast z_2) + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 (z_1 \ast x_2)
\]
then \(\exp(\beta_3)\) is equal to \(\text{RCRR}\) according to (9), but the other measures are not functions of \(\beta_3\) of this log-linear model. If we use a linear model

\[
\text{pr}(y = 1 | z, x) = \alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 (z_1 \ast z_2) + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 (z_1 \ast x_2)
\]
then \(\beta_3\) is equal to \(\text{DCRD}\) according to (13), but the other measures are not functions of \(\beta_3\) of this linear model. Furthermore, the ML estimations of the parameters in the log-linear and linear models do not converge even in the absence of the term for the product \(z_1 \ast x_2\). Robust methods could improve the convergence but need additional assumptions such as the Poisson distribution for the frequency of \(y = 1\) (see, for instance, [2]). Consequently, we are not able to obtain the ML estimates of \(\text{RCRR}\) and \(\text{DMRD}\) by these two models without additional assumptions.

If the term for the product \(z_1 \ast x_2\) in model (25) is omitted despite \(\theta_4 \neq 0\), then we obtain a misspecified model for the conditional risk \(\text{pr}(y = 1 | z, x)\)

\[
\log \left( \frac{\text{pr}(y = 1 | z, x)}{1 - \text{pr}(y = 1 | z, x)} \right) = \alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 (z_1 \ast z_2) + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3. \quad (26)
\]
This model may lead to biased estimate of \(\text{pr}(y = 1 | z, x)\) and thus biased estimates of \(\text{RCOR}\) and \(\text{RCRR}\) according to formulas (5) and (9). Furthermore, this model may lead to misspecification of \(\text{PR}(y = 1 | z)\) according to formula (14) and thus to biased estimates of \(\text{RMOR}\), \(\text{RMRR}\) and \(\text{DMRD}\) according to formulas (18), (21) and (24).

In the rest of this section, we are going to use models (25) and (26) separately to estimate \(\text{RCOR}\), \(\text{RCRR}\), \(\text{DMRD}\), \(\text{RMOR}\) and \(\text{RMRR}\) based on the data introduced earlier.

### 4.2 Conditional measures of interaction on population

First we describe the procedure of obtaining the ML estimates of the conditional measures \(\text{RCOR}\) and \(\text{RCRR}\) based on model (25). In model (25) replacing the parameters \(\pi = (\alpha, \beta_1, \beta_2, \beta_3, \theta_1, \theta_2, \theta_3, \theta_4)\) by the ML estimates \(\hat{\pi} = (\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)\) given by Table 2, we obtain the ML estimate \(\hat{\text{pr}}(y = 1 | z, x)\) of the conditional risk \(\text{pr}(y = 1 | z, x)\). In formulas (5) and (9) replacing \(\text{pr}(y = 1 | z, x)\) by \(\hat{\text{pr}}(y = 1 | z, x)\) and \(\text{pr}(x)\) by the proportion of \(x\) in the sample, we obtain the ML estimates \(\hat{\text{RCOR}} = 8.85\) and \(\hat{\text{RCRR}} = 1.60\), which are presented in Table 3.

To obtain interval estimate of \(\text{RCOR}\), we generate approximate distribution of the ML estimate \(\hat{\text{RCOR}}\). Normal distribution is poor approximation to the distribution of \(\hat{\text{RCOR}}\), because \(\text{RCOR}\) ranges from 0 to +\(\infty\). We have similar situations for the other measures of the interaction. On the other hand, the model parameters \(\pi = (\alpha, \beta_1, \beta_2, \beta_3, \theta_1, \theta_2, \theta_3, \theta_4)\) range from \(-\infty\) to +\(\infty\), so normal distribution is good approximation to the distribution of the ML estimate \(\hat{\pi} = (\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)\) (see, for instance, Lindsey, 1996). Let \(p\) be a random variable which follows the normal distribution \(N(\hat{\pi}, \hat{\Sigma})\), namely, \(p \sim N(\hat{\pi}, \hat{\Sigma})\), where \(\hat{\pi}\) and \(\hat{\Sigma}\) in \(N(\hat{\pi}, \hat{\Sigma})\) are given in Table 2. We are going to use this normal distribution to construct approximate distributions of \(\hat{\text{RCOR}}\) and \(\text{RCRR}\) in this subsection and those of \(\text{RMOR}\), \(\text{RMRR}\) and \(\text{DMRD}\) in the next subsection.

Now we describe the iteration procedure of generating approximate distribution of \(\text{RCOR}\). First, we draw \(p\) from the normal distribution \(N(\hat{\pi}, \hat{\Sigma})\) and replace \(\pi\) by \(p\) in model (25) to obtain a value of \(\hat{\text{pr}}(y = 1 | z, x)\). Second, we replace \(\text{pr}(y = 1 | z, x)\) by \(\hat{\text{pr}}(y = 1 | z, x)\) in (5) to obtain a value of \(\text{RCOR}\). We iterate the procedure, 1000 times in this article, to obtain an approximate
distribution of RCOR. The obtained approximate distribution is used to obtain the confidence interval of RCOR. Similarly, we use the normal distribution \( p \sim N(\hat{\pi}, \hat{\Sigma}) \) and model (25) and formula (9) to obtain an approximate distribution of RCRR and then the confidence interval of RCRR.

The obtained approximate distributions of RCOR and RCRR are shown in Figure 1. The 95% confidence interval of RCOR is (0.66, 127.04) while that of RCRR is (0.74, 4.60). They are also presented in Table 3, together with 50% confidence intervals of RCOR and RCRR.

Replacing model (25) by model (26) in the above procedure, we obtain the ML estimates and the confidence intervals of RCOR and RCRR based on model (26). The results are presented in Table 3.

### 4.3 Marginal measures of interaction on population

First we describe the procedure of obtaining the ML estimates of the marginal measures RMOR, RMRR and DMRD based on model (25). In model (25) replacing the parameters \( \pi = (\alpha, \beta_1, \beta_2, \beta_3, \theta_1, \theta_2, \theta_3, \theta_4) \) by the ML estimates \( \hat{\pi} = (\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4) \) given in Table 2, we obtain the ML estimate \( \hat{p}(y = 1 \mid z, x) \) of the conditional risk \( pr(y = 1 \mid z, x) \) in formula (14) replacing \( pr(y = 1 \mid z, x) \) by \( \hat{p}(y = 1 \mid z, x) \) and \( pr(x) \) by the proportion of \( x \) in the sample, we obtain the ML estimate \( \hat{PR}(y = 1 \mid z) \) of the population-adjusted risk \( PR(y = 1 \mid z) \). In formulas (18), (21) and (24) replacing \( PR(y = 1 \mid z) \) by \( \hat{PR}(y = 1 \mid z) \), we obtain the ML estimates \( \hat{RMOR} = 8.62, \hat{RMRR} = 1.58 \) and \( \hat{DMRD} = 0.34 \). These estimates are also presented in Table 3.

Now we describe the iteration procedure of generating approximate distributions of \( \hat{RMOR} \). First, we draw \( p \) from \( N(\hat{\pi}, \hat{\Sigma}) \) and replace \( \pi \) by \( p \) in model (25) to get \( \hat{p}(y = 1 \mid z, x) \). Second, we replace \( pr(y = 1 \mid z, x) \) by \( \hat{p}(y = 1 \mid z, x) \) in (14) to get \( \hat{PR}(y = 1 \mid z) \). Finally, we replace \( PR(y = 1 \mid z) \) by \( \hat{PR}(y = 1 \mid z) \) in formula (18) to get \( RMOR \). We iterate the procedure, 1000 times in this article, to get an approximate distribution of \( RMOR \). The approximate distribution is used to get the confidence interval of \( RMOR \). Similarly, we use the normal distribution \( p \sim N(\hat{\pi}, \hat{\Sigma}) \) and formulas (25), (14) and (21) to get an approximate distribution of \( RMRR \) and then the confidence interval of \( RMRR \), and the normal distribution and formulas (25), (14) and (24) to get an approximate distribution of \( DMRD \) and then the confidence interval of \( DMRD \).

The obtained approximate distributions of RMOR, RMRR and DMRD are shown in Figure 1. The 95% confidence interval of RMOR is (0.72, 88.51), that of RMRR is (0.77, 3.40), and that of DMRD is (−0.10, 0.72). These confidence intervals are also presented in Table 3, together with the 50% confidence intervals of RMOR and RMRR and DMRD.

Replacing model (25) by model (26) in the above procedure, we get the ML estimates and the confidence intervals of RMOR, RMRR and DMRD based on model (26). The results are given in Table 3.

### 4.4 Results

In practical researches, one usually uses model (26), which omits the term for the product \( z_1 \ast x_2 \), to measure and estimate the interaction as described in the introduction. In this case, one can only have RCOR. Because the outcome \( y = 1 \) is common in this study, one cannot convert RCOR into RCRR or DMRD.
Despite the common outcome $y = 1$ and the term for the product $z_1 \times x_2$, we have used one logistic model (25) to estimate all the measures of interaction, RCOR, RCRR, RMOR, RMRR and DMRD. Figure 1 shows the approximate distributions of the ML estimates $\text{RCOR}$, $\text{RCRR}$, $\text{RMOR}$, $\text{RMRR}$ and $\text{DMRD}$ based on model (25). Table 3 summarizes the ML estimates and the 50% and 95% confidence intervals for the five measures RCOR, RCRR, RMOR, RMRR and DMRD based on model (25) in comparison to the mis-specified model (26).

By comparing the second column with the third column in Table 3, we see that model (26) causes considerable biases to all the five measures. These biases might be exacerbated if more terms for complex products between exposures and covariates were needed to model $\Pr(y = 1 \mid z, x)$ as typical of a large sample. Our method can avoid these biases by being able to use complex models.

Model (25) fits the data best and thus provides the most efficient point and interval estimates for these measures of interaction. A clear advantage of using one model in one study to estimate different measures of interaction is that one can avoid different assumptions behind different models, which may complicate interpretations of these measures.

Despite somewhat wide confidence intervals, all the five measures of interaction based on model (25) suggest that the therapy would have higher efficacy of eradicating *Helicobacter pylori* if children lived in the rural environment than if children lived in the urban environment. This finding suggests that urbanization may have contributed to the low efficacy of the therapies containing antibiotics among children in developing countries. To our knowledge, little is seen in the literature about how the environment interacts with therapies on eradication of the bacterium [16, 17].

5. Discussion and conclusions

An underlying interaction between exposures on a certain population has different aspects described by different measures. Depending on their specific research subjects, researchers need to know one or several aspects of the interaction. In this article, we have used one model in one study to obtain point and interval estimates for five different measures. The outcome can be common. The model may contain terms for the products between exposures and covariates. The estimation is based on maximum likelihood. The method can be implemented by using any software that generates normal distribution.

In the literature, two methods are available to obtain confidence intervals of various measures of exposure effect, i.e. the normal approximation method and the bootstrap method [1, 6, 8, 13], but which seem less used to obtain confidence intervals for measures of interaction. The normal approximation method is to derive approximate variance of the ML estimate of a measure of exposure effect by the delta method and then use the variance to obtain normal approximation confidence interval of the measure of exposure effect. To obtain normal approximation confidence interval of a measure of interaction, one needs to derive approximate variance of the ML estimate of the measure of interaction, but the derivation is tedious particularly for complex models.

The bootstrap method is to generate bootstrap samples by parametric or non-parametric bootstrap method and then use the bootstrap samples to obtain the bootstrap distribution of a measure of exposure effect and then the bootstrap confidence interval of the measure of exposure effect. However, it is highly difficult to correct the finite-sample bias arising from the bootstrap sampling with dichotomous outcome [3, 4, 9], and the difficulty would be exacerbated for a measure of interaction.
In our method, we generate approximate distribution of the ML estimate of a measure of interaction and use the distribution to obtain the confidence interval for the measure of interaction. This method provides a simple but reliable approach to interval estimation of the measure of interaction. Although it is beyond the scope of this article, it is of interest to derive the corrected bootstrap confidence interval and the normal approximation confidence interval and compare them with the maximum-likelihood-based confidence interval obtained in this article.

References

[1] P.C. Austin, *Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model*, Journal of Clinical Epidemiology 63 (2010), pp. 2–6.
[2] G.S. Bieler, G.G. Brown, R.L. Williams and D.J. Brogan, *Estimating Model-Adjusted Risks, Risk Differences, and Risk Ratios From Complex Survey Data*, American Journal of Epidemiology 171 (2010), pp. 618–623.
[3] J. Carpenter and J. Bithell, *Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians*, Statistics in Medicine 19 (2000), pp. 1141–1164.
[4] A.C. Davison and D.V. Hinkley, *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge, 1997
[5] M.H. Gail, S. Wieand and S. Piantadosi, *Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates*, Biometrika 71 (1984), pp. 431-444.
[6] B.I. Graubard and E.L. Korn, *Predictive Margins with Survey Data*, Biometrics 55 (1999), pp. 652–659.
[7] S. Greenland, *Interactions in Epidemiology: Relevance, Identification and Estimation*, Epidemiology 20 (2009), pp. 14–17.
[8] S. Greenland, *Model-based Estimation of Relative Risk and Other Epidemiologic measures in Studies of Common Outcomes and Case-Control Studies*, American Journal of Epidemiology 160 (2004a), pp. 301–305.
[9] S. Greenland, *Interval estimation by simulation as an alternative to and extension of confidence intervals*, International Journal of Epidemiology 33 (2004b), pp. 1389–1397.
[10] S. Greenland, J.M. Robins and J. Pearl, *Confounding and collapsibility in causal inference*, Statistical Science 14 (1999), pp. 29–46.
[11] S. Greenland and J.M. Robins, *Identifiability, exchangeability, and epidemiological confounding*, International Journal of Epidemiology 15 (1986), pp. 413–419.
[12] J. Guo and Z. Geng, *Collapsibility of logistic regression coefficients*, Journal of Royal Statistical Society B 57 (1995), pp. 63–267.
[13] P.W. Lane and J.A. Nelder, *Analysis of Covariance and Standardization as Instance of Prediction*, Biometrics 36 (1982), pp. 613–621.
[14] Y. Lee and J.A. Nelder, *Conditional and marginal models: another view*, Statistical Science 19 (2004), pp. 219–228.
[15] J.K. Lindsey, *Parametric Statistical Inference*, Oxford, Clarendon Press, 1996
[16] P. Malferttheiner, F. Megraud, C. O'Morain, F. Bazzoli, E. El-Omar and D. Graham, *Current concepts in the management of Helicobacter pylori infection: the Maastricht III Consensus Report*, Gut 56 (2007), pp. 772–781.
[17] F. Megraud, *H. pylori antibiotic resistance: prevalence, importance, and advances in testing*, Gut 53 (2004), pp. 1374—1384.
[18] T.V.H. Nguyen, C. Bengtsson, G.K. Nguyen, T.T.H. Hoang, D.C. Phung, M. Sorberg and M. Granstrom, *Evaluation of Two Triple-Therapy Regiments with Metronidazole or*
Clarithromycin for the Eradication of H. pylori Infection in Vietnamese Children, a Randomized, Double Blind Clinical Trial, Helicobacter 13 (2008), pp. 550--556.

[19] P.R. Rosenbaum, *Observational studies*, New York, Springer, 1995.
[20] P.R. Rosenbaum and D.B. Rubin, *The central role of the propensity score in observational studies for causal effects*, Biometrika 70 (1983), 41--55.
[21] K.J. Rothman, S. Greenland and T.L. Lash, *Modern Epidemiology (3rd edition)*, Lippincott Williams & Wilkins, Philadelphia, 2008
[22] K.J. Rothman, S. Greenland and A.M. Walker, *Concepts of interaction*, American Journal of Epidemiology 112 (1980), pp. 467--470.
[23] D.B. Rubin, X. Wang, L. Yin and E. Zell, *Estimating the Effect of Treating Hospital Type on Cancer Survival in Sweden Using Principal Stratification* in The HANDBOOK OF APPLIED BAYESIAN ANALYSIS (Eds. T. O’Hagan and M. West), Oxford University Press, Oxford, 2010.
[24] T.J. VanderWeele, *On the Distinction Between Interaction and Effect Modification*, American Journal of Epidemiology 20 (2009), pp. 863--871.
Table 1 Successful eradications / total patients stratified by the covariates $x_1$ (age), $x_2$ (gender), and $x_3$ (antibiotic resistance) for the treatment $(z_1, z_2)$ (therapy, environment).

| Covariates $x_1$ $x_2$ $x_3$ | Treatment $(z_1, z_2)$ $(0, 0)$ $(0, 1)$ $(1, 0)$ $(1, 1)$ |
|-------------------------------|------------------|----------------|----------------|----------------|
| 0 0 0                         | 3/4              | 6/7            | 0/3            | 8/8            |
| 0 0 1                         | 3/4              | 1/1            | 1/1            |                |
| 0 1 0                         | 1/5              | 3/4            | 2/3            | 3/3            |
| 0 1 1                         | 2/4              | 1/2            | 1/1            | 1/1            |
| 1 0 0                         | 1/2              | 1/1            | 3/8            | 6/6            |
| 1 0 1                         | 3/3              | 0/1            | 1/1            | 3/3            |
| 1 1 0                         | 2/7              | 0/2            | 5/8            | 0/2            |
| 1 1 1                         | 1/2              | 0/3            | 6/9            | 1/1            |
Table 2 ML estimates and its approximate covariance matrix for parameters of the model (25)*

| Parameters | $\alpha$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ |
|------------|---------|---------|---------|---------|---------|---------|---------|---------|
| Estimates  | 1.19    | -0.87   | 0.10    | 2.18    | -0.57   | -1.82   | 0.55    | 1.96    |

Covariance matrix

|       | $\alpha$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ |
|-------|---------|---------|---------|---------|---------|---------|---------|---------|
| $\alpha$ | 0.64    | -0.53   | -0.32   | 0.29    | -0.12   | -0.42   | -0.13   | 0.47    |
| $\beta_1$ | -0.53   | 1.06    | 0.31    | -0.67   | -0.14   | 0.45    | 0.10    | -0.89   |
| $\beta_2$ | -0.32   | 0.31    | 0.72    | -0.71   | 0.00    | 0.05    | 0.06    | -0.07   |
| $\beta_3$ | 0.29    | -0.67   | -0.71   | 1.86    | 0.05    | -0.05   | -0.04   | 0.32    |
| $\theta_1$ | -0.12   | -0.14   | 0.00    | 0.05    | 0.37    | -0.04   | -0.05   | 0.03    |
| $\theta_2$ | -0.42   | 0.45    | 0.05    | -0.05   | -0.04   | 0.69    | -0.01   | -0.69   |
| $\theta_3$ | -0.13   | 0.10    | 0.06    | -0.04   | -0.05   | -0.01   | 0.39    | -0.10   |
| $\theta_4$ | 0.47    | -0.89   | -0.07   | 0.32    | 0.03    | -0.69   | -0.10   | 1.40    |

*The covariance matrix is obtained with adjustment of over-dispersion in the framework of the ML estimation of a mis-specified model or the generalized estimating equation.
Table 3 ML estimates and confidence intervals for the five measures of the interaction describing how efficacy of the therapy on eradication of *Helicobacter pylori* would increase if children lived in the rural environment of Vietnam, based on models (25) and (26).

| Interaction | ML estimates | 50 % and 95 % confidence intervals |
|-------------|--------------|-----------------------------------|
|             | Model (25)   | Model (26)                        |
| Reservoir  | RCOR         | 8.85                              | (0.66, 3.61, 21.95, 127.04) | (0.40, 2.18, 13.89, 72.00) |
| Effort      | RCRR         | 1.60                              | (0.74, 1.22, 2.27, 4.60)    | (0.60, 1.07, 2.05, 4.18)    |
| Impact     | RMOR         | 8.62                              | (0.72, 3.50, 15.93, 88.51)  | (0.42, 2.04, 11.32, 39.45)  |
| Response   | RMRR         | 1.58                              | (0.77, 1.23, 1.90, 3.40)    | (0.63, 1.06, 1.80, 3.10)    |
| Ratio      | RMRD         | 0.34                              | (−0.10, 0.18, 0.45, 0.72)   | (−0.21, 0.09, 0.39, 0.65)   |
**Figure 1** Approximate distributions of the ML estimates -- $\hat{R}_{COR}$, $\hat{R}_{CRR}$, $\hat{R}_{MOR}$, $\hat{R}_{MRR}$ and $\hat{D}_{MRD}$ -- for the five measures of the interaction between the therapy and the environment on the eradication of *Helicobacter pylori* among Vietnamese children.