Multilevel regression modelling to investigate variation in disease prevalence across locations

Gudrun Weinmayr*[1], Jens Dreyhaupt[1], Andrea Jaensch[1], Francesco Forastiere[2] and David P Strachan[3]

1 Institute of Epidemiology and Medical Biometry, Ulm University, Ulm, Germany

2 Department of Epidemiology, Lazio Regional Health Service, Rome, Italy

3 Population Health Research Institute, St George's, University of London, London, United Kingdom

*Corresponding author:

Gudrun Weinmayr, Institute of Epidemiology and Medical Biometry, Ulm University, Helmholtzstr.22, 89081 Ulm, Germany. E-mail: gudrun.weinmayr@uni-ulm.de

Abstract:

In this article, we show how to investigate the role of individual (personal) risk factors on outcome prevalence in multicentre studies with multilevel modelling. The variation in outcome prevalence is modelled by introducing a random intercept. In the next step, the empty model is compared with the model containing the risk factor(s). Because the outcome is dichotomous, this comparison can only carried out after having rescaled the models' parameter values to the variance of an underlying continuous variable. We illustrate this approach with data from Phase Two of the International Study of Asthma and Allergies in Childhood (ISAAC) and provide a corresponding Stata do-file.

Key words: prevalence variation, risk factor, multicentre study, multilevel regression, random intercept

Introduction:

Many diseases vary widely in prevalence worldwide and a most plausible explanation is the influence of environmental risk factors. Environmental risk factors vary also widely in prevalence over the world and often provide exposure contrasts between populations where within population contrasts are low [1]. Investigation of differences between diverse populations has given important insights for instance regarding cardiovascular risk factors [2] and asthma and allergies in children [3].

However, simple ecological comparisons, using aggregated data only, do not take into account associations at the individual level and are potentially prone to ecological fallacies. A multilevel approach can take into account the contribution of individual risk factors on disease prevalence in different locations (e.g. countries), thus shedding light on potential ecological fallacies.

The more usual focus on the association between disease and risk factors within populations ignores the broader issue of the importance of these factors in determining the burden of disease in whole populations and therefore the potential for prevention at a community level.

The multilevel approach can be used to investigate both the variation between locations and the determinants of this variation. These will be risk factors that have an association with the disease at the individual level but also vary in their prevalence between locations or study centres.

Although the multilevel framework has the potential of investigating both the individual and population level determinants of disease variation it seems to have been used very little in epidemiology. So far studies investigating the differences between locations come rather from the social sciences such as the classic example by Lee & Bryk on mathematics achievement in different schools [4] or newer work e.g. investigating children's wellbeing using indicators [5]. In the health sector, Wainwright & Surtees [6] give an example of analysing functional health scores.

In these cases the outcome investigated is generally continuous and linear multilevel or "hierarchical linear models" can be used. In such models the total variance can be easily decomposed in three variance components: 1) the explained variance that can be attributed by variation in a predictor variable (e.g. risk factor), 2) the unexplained variance (or residual variance) at the individual level and 3) the unexplained variance at the centre level. The latter is equivalent to the variation of the random intercept. The advantage of linear models is that models with different number of risk factors can be easily compared regarding the change of their parameters and therefore also of the variation between locations. Hence these models allow a direct assessment of the change in variation when a risk factor is introduced into the model.

In epidemiology, however, outcomes are very frequently binary and the corresponding multilevel logistic regression models do not allow such a simple comparison between models, although this sometimes seems to have been ignored (for example [7–9]). For a binary variable, the residual variance at the individual level is determined by the binomial distribution and has a fixed value of $\pi^2/3$ in case of the logit model. As a consequence of this fixed individual residual variance, all the other parameters in the model (including the variance of the random intercept) change to accommodate the change introduced by an individual-level explanatory variable. To allow any meaningful interpretation the relative changes of the parameters have to be assessed [10].

Bauer[11] proposed a scaling procedure to achieve comparable estimates between different probit or logistic regression models, based on the assumption of an underlying continuously distributed variable. Fielding [12] had earlier introduced the same concept and a similar scaling procedure for mutinomial regression. . A scaling procedure which is mathematically equivalent to Bauer's, published later by Hox [13] has been implemented as a Stata ado file by Enzmann and Kohler [14]. This compares any model with risk factors to a null model without individual level risk factors, whereas Bauer's method can compare different risk factor models directly.

The consequences of ignoring the scaling procedure are biased estimates of changes in variance components, with bias possible in either direction. Fielding ([12]) used real data on mathematical test scores in schoolchildren, modelled by multinomial regression, to investigate changes in the variance of the underlying latent variable, representing test performance, between 114 schools, after adjustment for a range of individual (child-level) covariates. The change in school-level variance was overestimated by as much as 118% and underestimated by as much as 74%, when different risk factor combinations were introduced into the model.

Not taking into account the scaling effect leads to incorrect values not only of the level-two variance but also of the regression coefficients. E.g. Bauer[11] reports in his simulated data a marked increase in the beta coefficient of the level-1 variable after introduction of an additional level-1 variable (from 1.51 to 2.65) which after scaling of the results turns out to be no change (from 1.01 to 1.02). In the same example, the variance of the random intercept showed no change when not scaled, in contrast to a decrease from 0.16 to 0.06 when scaled.

While the estimates of the random intercept variance are routinely reported with their variance, up to date no routine exists to estimate the uncertainty in the estimated changes of the random intercept variance. According to a suggestion by Bauer [11], we apply, to our knowledge for the first time, bootstrap methods to estimate the related confidence intervals.

In summary, without a scaling procedure it is not possible to quantify the variance in the random intercept that reflects the prevalence variation of any binary outcome, and therefore not possible to assess the influence of any risk factor on this prevalence variation. Therefore, in this tutorial we show how to apply this approach to the investigation of variation in disease prevalence across locations using the data of the Phase Two of the International Study of Asthma and Allergies in Childhood (ISAAC). A corresponding Stata Macro allowing the direct comparison of several different multivariate models with different risk factors has been developed based on a SAS-Macro developed by Bauer (personal communication).

The use of multilevel models for modelling prevalence variation

In an ordinary one-level regression model the assumption is that all individuals, even if from different centres, belong to one common population. In a multi-level model, we consider that there may be genuine differences between centre populations, which are themselves a sample from a superpopulation. The corresponding variation in the outcome is accounted for by the random intercept in the multilevel model and therefore is equivalent to modelling the prevalence of centres – indeed in logistic regression the intercepts are the centre-level prevalence logodds.

To model the random intercept, we introduce a random effect term at the centre level (level 2). Note that in contrast to a linear model, the formula for the logistic model, does not contain a random effect term (i.e. error term) at the individual level (level 1) because this error term is determined by the binomial distribution, in the case of a logit link it is $\frac{\pi^2}{3}$.

Therefore, instead of the usual equation for single-level logistic regression models

Logit(p_i) = b_0 + $\mathbf{bX_i}$

where $\mathbf{X_i}$ is (are) the independent variable(s) (e.g. risk factor(s)) of individual i

the corresponding equation for a two level model writes

Logit(p_{ij}) = γ_{00} + $\mathbf{bX_{ij}}$ + u_j

where $\mathbf{X_{ij}}$ is (are) the independent variable(s) (e.g. risk factor(s)) of individual i in centre j and u_j denotes the level 2 (i.e. centre-level) random intercept.

u_j follows a random distribution which has a variance denoted here as tau^2. This variance is equivalent to the variance between centres in the random intercept and therefore to the variance around the grand mean of the prevalence logodds (γ_{00}) as reflected in the full dataset.

Here we deal with models where the centres have the same slope (representing a fixed effect of the risk factor on disease logodds) and vary only in the intercept (Figure 1). Thus, we use a multilevel model with a random intercept i.e. there is only a second level variation

term (error term) for the intercept. Multilevel models can also incorporate random variation of the slope which, however, is beyond the scope of the current article. Introduction of random slopes poses additional problems of analysis and interpretation of the distribution of random intercepts, because the variance of these intercepts will vary according to the choice of origin (zero exposure) for the explanatory variables with non-fixed slopes.

As in other regression models explanatory variables can be introduced which will lead to a change in the variance explained by the model. In the case of a continuous outcome measure, and a multi-level linear regression, we have two model terms for the variation, one for level 1 (individual subjects) and one for level 2 (study centres or locations) and therefore we can investigate specifically the change of the residual variance between individuals and the change of the residual variance between centres directly from the model output. This means, if we want to assess the amount of between centre variance that can be explained by a risk factor, we can look directly at the change in the variance of the centre-level random intercepts u_j after introduction of an explanatory variable.

This is because the residual (unexplained) variances at both the centre level and the individual level are reduced by introducing explanatory factors to a linear model.

In contrast, in a multilevel logistic model the variance at the individual level is defined by the distribution of the binary variable and remains therefore constant (e.g. at $\frac{\pi^2}{3}$ in case of the logit link function) in all models. Introduction of any risk factor into the model can therefore not change the constant unexplained residual variance at the individual level. Instead, all the beta-coefficients and the variance components in the model change reflecting the change of the residual variance by the newly introduced variable. In case of the investigation of the variance between centres we are interested in the change in tau^2 in relation to the total variance that consists of the explained variance and the residual variance at individual level and the residual variance at the centre level tau^2. Snijders and Bosker [10] therefore propose a formula which allows a quantification of the proportion of the respective variance component

relative to the total variance (decomposition of the total variance into the three components: explained variance, residual variance at level 2 and residual variance at level 1).

In a further step Bauer [11] developed a scaling procedure based on the concept of an underlying continuous latent variable. Here the binary variable is defined by an underlying latent continuous variable y*. The binary outcome y=0 if y* is below a threshold value T and y=1 if the value is equal or above T. The approach proposed by Bauer [11] rescales all model parameters to the constant total variance of this underlying continuous latent variable y*. In the specific case of a model with a random intercept but fixed slopes, his general formula reduces to

$s = a/(\boldsymbol{b}' \sum \boldsymbol{X} \, \boldsymbol{b} + tau^2 + \sigma_r^2)$ (equation (eq.) 1) where

    s is the scaling factor,

    a is the specific value of the variance of the underlying response variable,

    $\boldsymbol{b}$ is the vector of the regression coefficients of the fixed effects, and $\boldsymbol{b}'$ its transpose

    $\sum \boldsymbol{X}$ the covariance matrix of the risk factors enhanced by additional 0-cells to match in dimension $\boldsymbol{b}$ and $\boldsymbol{b}'$,

    σ_r^2 is the variance of the individual level random effect.

Note that a and σ_r^2 are identical, e.g. $\frac{\pi^2}{3}$ for logistic regression, or 1 for a probit regression. Thus conceptually the scaling factor is the ratio of the individual-level residual (binomial sampling) variance ($\frac{\pi^2}{3}$ for logistic regression) to the sum of the variance explained at the individual level ($\boldsymbol{b}' \sum \boldsymbol{X} \, \boldsymbol{b}$), the unexplained centre-level variation (tau^2) and the unexplained (sampling) variation at the individual level ($\frac{\pi^2}{3}$ for logistic regression).

In the case of a model without risk factors, the formula for the scaling factor reduces further to $s = a/(tau^2 + \sigma_r^2)$ (eq. 2).

Only parameters rescaled with this scaling factor can be compared directly between models. Therefore, this scaling procedure is necessary when assessing the reduction in between-

centre variance after introducing an individual-level risk factor. Note that the scaling factor for the effect estimates (**b**) is the square root of the above scaling factor for the variance.

<u>Introduction into the worked example: The ISAAC Phase Two data set</u>

The data set used in this example comes from the Phase Two of ISAAC which is a multicentre cross-sectional study that was carried out in 30 centres in 21 countries and investigated children on average 9-11 years old [15]. Outcome data derive from standardized questionnaires on symptoms of asthma and from the results of skin prick tests to six aeroallergens [16].

The main symptom of asthma used in analysis is "wheeze in the past year" which is a binary variable "yes"-"no". A skin prick test was rated as positive if the wheal size in response to any of the six allergens was greater than 3 mm, after subtraction of the negative control. The variables from the risk factor questionnaire we use in this example are mostly dichotomous, but also categorical or continuous variables may be introduced into the model. All questionnaires were filled in by the care taker, usually parents. The detailed questionnaires are available on http://isaac.auckland.ac.nz/phases/phasetwo/phasetwomodules.pdf.

In this article we use a subset of centres that had investigated risk factors in the full sample (as opposed to a subsample stratified by wheeze) and information on the respective risk factor and outcome. Only children with information on the risk factors used in the examples are included in these restricted data sets: 27 719 children from 24 centres contribute to the example examining skin prick test positivity and number of siblings, and 34 809 children from 24 centres and contribute to the example investigating wheeze and four selected risk factors.

These four selected risk factors were: Mother smoked during the first year of life of the child (yes/no), the child has had a whooping cough infection (yes/no), the child had been breastfed (yes/no) and fish consumption at least once a week (yes/no).

Examples of analyses on the Phase Two data

We explore two examples, the numbers of siblings in relation to atopy and the influence of four risk factors on asthma. The number of siblings has been shown to be an important factor influencing the occurrence of allergic disease[17] and we chose therefore to investigate its relation with an objective marker of allergy i.e. skin prick test positivity. The number of siblings is furthermore very variable in the international context with higher mean numbers in less affluent centres. On theoretical grounds we would expect an individual risk factor that has a) a reasonable strong effect size and b) a high variability between populations to have a notable effect on prevalence differences worldwide.

After a first exploration of the Phase Two data sets we chose the risk factors for our wheeze example to reflect different directions of association (harmful and protective) on the individual level and the ecological level. On the individual level, maternal smoking and whooping cough infection represent harmful factor and breastfeeding and fish consumption the protective factors. On the ecological level, maternal smoking and breastfeeding have positive correlation with wheeze prevalence, whereas whooping cough infection and fish consumption have a negative correlation. As we will see both the individual and ecological level matter for the direction of the overall change in variation, highlighting that ecological analyses on their own may be misleading in this kind of analysis. The prevalence of all four risk factors varies strongly in the international context.

a) Example 1: number of siblings and atopy

In a first example we consider how much of the international variation in the prevalence of atopy, defined as a positive skin prick test, can be explained by the reported number of siblings and birth order as reflected by the number of older siblings.

To evaluate the change in the rescaled tau^2 when introducing an explanatory variable, we first calculate an empty model without the risk factor to determine the variation in wheeze prevalence as reflected by tau^2 which is 0.6061. This entity, we call tau_0^2.

We use eq. 2 to calculate the scaling factor. In our case of a logistic regression model for a binary outcome, a and σ_r^2 in the equation are $\frac{\pi^2}{3}$. We obtain a scaling factor of 0.844 and therefore a rescaled tau_0^2 of 0.5115.

We then introduce the number of siblings (variable num_sibs) into the model (see Figure 2) and obtain a tau^2 of 0.5657 which is (0.7521^2) i.e. the sd(_cons) from Figure 2 squared.

For the formula we now need to calculate the term $\boldsymbol{b}' \sum \boldsymbol{X} \, \boldsymbol{b}$ (see eq.1).

$\boldsymbol{b}$, the vector of the regression coefficients, starting with the intercept, is here (-1.296155 -0.0542936) see Figure 2, and therefore its transpose is $\begin{pmatrix} -1.296155 \\ -0.0542936 \end{pmatrix}$. Note that the order of the regression coefficients in Stata is actually in the inverse order of that needed in the formula (e.g. intercept in the last instead of the first position).

$\sum \boldsymbol{X}$, is obtained in Stata with "correlate num_sibs, cov" and contains the covariance matrix, in the single risk factor case simply the variance of that variable (3.19307), and is then enhanced with additional 0s to obtain

$$\begin{pmatrix} 0 & 0 \\ 0 & 3.19307 \end{pmatrix}$$

The whole formula for the scaling factor according to (eq. 1) therefore is

$$\frac{\frac{\pi^2}{3}}{\left(\begin{pmatrix} -1.296155 \\ -0.0542936 \end{pmatrix} \times \begin{pmatrix} 0 & 0 \\ 0 & 3.19307 \end{pmatrix} \times \left(-1.296155 \quad -.0542936 \right) \right) + .56572603 + \frac{\pi^2}{3}}$$

which results in a scaling factor of 0.8511 and multiplying it with tau^2 we obtain the rescaled

tau^2 of 0.4815. This results in a relative reduction respective to the null-model of 5.91%.

We next introduce a second variable, the number of older siblings (oldersibs) into the model.

As we have already the number of all siblings in the model, this variable reflects mainly birth

order and informs us whether having older sibling has a stronger effect than having younger

siblings.

b, the vector of the regression coefficients, starting with the intercept followed by the first and

then the second explanatory variable, is now (-1.295132 -0.0576607 0.005064) see Figure

3, and therefore its transpose is $\begin{pmatrix} -1.295132 \\ -.0576607 \\ 0.005064 \end{pmatrix}$.

$\sum X$ is obtained in Stata with "correlate num_sibs oldersibs, cov" and contains the variance of

num_sibs (3.1931) and the variance of oldersibs (1.8923), and the covariance between the

two variables (1.9859) (Figure 4). This is then enhanced with additional 0s to obtain

$$
\begin{matrix}
0 & 0 & 0 \\
0 & 3.19307 & 1.9859 \\
0 & 1.9859 & 1.89234
\end{matrix}
$$

The whole formula for the scaling factor according to (eq. 1) therefore is

$$
\frac{\frac{\pi^2}{3}}{\left(\left(\begin{smallmatrix} -1.295132 \\ -0.0576607 \\ 0.005064 \end{smallmatrix} \right) \text{ x } \begin{pmatrix} 0 & 0 & 0 \\ 0 & 3.19307 & 1.9859 \\ 0 & 1.9859 & 1.89234 \end{pmatrix} \text{ x } \left(-1.295132 \quad -0.0576607 \quad 0.005064 \right) \right) + 0.56534304 + \frac{\pi^2}{3}}
$$

which results in a scaling factor of 0.8513 and multiplying it with tau^2 we obtain the rescaled

tau^2 of 0.4813. This now results in a relative reduction respective to the null-model of 5.97%.

This is very similar to the figure of 5.91% obtained from the single risk-factor model with total

number of siblings. This reflects both the weakness of the older siblings effect at the

individual level, independent of total siblings, and the fact that mean birth order does not vary

greatly between study centres independently of total sibship size, whereas there is much greater variation in total sibship size.

b) Example 2: Influence of several risk factors on wheeze in the past year

In the second example we investigate several risk factors for their influence on the variation of wheeze prevalence.

We start by investigating each risk factor separately for the change in tau after this risk factor has been introduced into the random intercept model without any explanatory variables ("empty model"). We then start building a multivariate model introducing the explanatory factors one-by-one.

The influence of the tested risk factors when introduced as only risk factors in the model on the change in tau^2 is shown in Table 1. The change in tau^2 is presented as the relative decrease (in %) in relation to the tau^2 of the empty model and hence a negative sign indicates that the risk factor actually increases tau^2.

The variable "mother smoking in the first year of life of the child" has a moderate association with wheeze at the individual level with an odds ratio (OR) of 1.2 which is statistically significant. This risk factor is also quite variable in its prevalence across the centres (Table 1). The introduction of this risk factor results in a relative reduction in the variability between centres, as reflected by tau^2, of 7.1%. This risk factor has a positive ecological correlation with wheeze and hence the tau^2 is decreased.

Whooping cough has a moderately higher and also statistically significant effect size and a somewhat lower variation in prevalence between centres. However, this factor entrains an increase in tau. This seems at first glance counterintuitive because one might think that introduction of any predictive factor will decrease the variance. The increase in tau^2 is a consequence of the negative ecological correlation between the wheeze prevalence and the

whooping cough infection prevalence – in contrast to the prevalence of mothers smoking where we had a positive correlation with wheeze prevalence at the centre level (Figure 5)

We also observe an increase in tau^2 for breastfeeding, where the ecological correlation is positive but the association at individual level is inverse i.e. smaller than one. On the other hand, for fish consumption, where both the odds ratio as well as the correlation are negative we observe again a decrease in tau^2.

The final model (Table 2), incorporating individual level variables that decrease and increase the tau^2 results in an overall decrease of 1.8%.

The precision of our estimates of the change in tau^2, is not obtainable directly from model outputs, so we have applied bootstrapping to determine its 95%-confidence interval (CI). As we are looking at a centre-level parameter, bootstrapping should not sample individual children but needs to sample centres, i.e. a block sampling procedure has to be performed. Performing this procedure 2000 times is sufficient to derive 95% CI [18]. For carrying out the block sampling procedure in our worked example, complete centres were randomly selected with replacement from the whole study population until the number of centres reached 24 in each bootstrap sample. Because of different numbers of children in each centre, the number of children varied between the bootstrap samples. The empty model and the full model were fitted to each bootstrap sample and afterwards the results were rescaled. A simple 95% bootstrap confidence interval was obtained using the 2.5% and the 97.5% percentile as confidence limits. For the example with skin prick test and siblings, we obtain thus a relative reduction of unexplained between-centre variance of 6% with the CI ranging from -0.5% to 14%. For the wheeze example with its small decrease of unexplained between-centre variance of 1.8% we obtain a CI of -7% to 8.8%.

In the appendix we present the code of a Stata macro that runs the multilevel analyses and calculation for models with as many risk factors as desired. This procedure can easily be adapted to probit regression (as used by Bauer [11]) by replacing $\pi^2/3$ by 1. In our data set, the

results and scaling factors obtained by probit regression were very similar to the results obtained by logistic regression.

Our Stata-Macro can be adapted to accommodate sampling weights which are necessary often in survey data or when analysing stratified subsamples (see code in the appendix). For example, the ISAAC Phase Two data consists partly of subsamples stratified by wheeze i.e. 100 wheezers and non-wheezers were sampled in some centres for certain modules [15]. In this case, we use gllamm to incorporate the appropriate "pweight". Note, however that certain conditions have to apply to get reliable unbiased results from this weighted multilevel approach using a so-called *pseudolikelihood*. For details we refer the reader to Rabe-Hesketh & Skrondal [19] and Rabe-Hesketh & Skrondal [20] on page 572. Note also that stratum sampling weights remain necessary even if the outcome variable is the one that was used to define the stratified subsample. This is in contrast to the more conventional analyses of the effects of risk factors on the outcome itself, where no weighting is needed because the analyses then correspond to a classical case-control analysis. However in our analysis, we are interested in the intercept and its variation and not in the effect estimates. Therefore not weighting the children in the subsample to the frequencies in the full sample will result in falsely elevated intercepts in the centres concerned – just as we cannot interpret the intercept as baseline prevalence in a case-control study. The consequence is a grossly false result for the overall intercept and its variation tau^2. Note that the use of gllamm with weights necessitates markedly more time than doing an unweighted analysis with xtlogit. Advice for speeding up the estimation procedure with gllamm is given in Rabe-Hesketh & Skrondal [20] on page 542. Even so, in the complete Phase Two data set with more than 50 000 observations and 30 centres, one multivariate model can run for several hours.

General Discussion

Our application of Bauer's method [11] to rescale the model parameters to an underlying continuous outcome offers a useful way to investigate the role of individual-level risk factors

in explaining variations in disease prevalence between populations. This method allows standardizing the interpretation of estimates of the tau^2, i.e. of the unexplained non-sampling variance of the random intercept. We therefore can estimate the proportion of the variance between the centres that can be explained by individual-level risk factors that are included in the model. In our example of real data, we could show e.g. that the unexplained variance in wheeze prevalence between centres is estimated to be reduced by 7% when maternal smoking in pregnancy was taken into account.

The between-centre variance tau^2 is comparatively high in our data set with unrescaled values of 0.52 (skin prick test example) and 0.29 (wheeze example). The relatively high tau^2 in our dataset corresponds to a prevalence range of 1.6% to 43.6% for the skin prick test example and of 3.4% to 26.6% for the wheeze example. Other studies have found e.g. a tau^2 of 0.14 for a study of reported asthma diagnosis in children in 287 neighbourhoods and 45 000 individuals [21]. Other studies on diverse outcomes found an even lower tau^2 [8,22,23]. We are not aware of any study reporting corresponding figures for a binary outcome in an international multi-centre study.

Our modelling with real world data showed that the incorporation of individual level variables, in contrast to centre level variables, can either decrease or increase the unexplained between-centre variance (tau^2). Therefore multivariate modelling can lead to very little net change when factors increasing and factors decreasing the variance are simultaneously incorporated into the model. A special case consists in introducing variables that reflect very similar exposures like "smoking of the mother in the first year of life of the child" or "smoking of the mother during pregnancy". These variables are highly correlated but as we are not interested in their effect estimates but in the random intercept, the collinearity of these two variables is not a problem[24]. Indeed the standard error (SE) of the tau^2 remains very similar between the models containing one or both of the risk factors. However, in general the additional variables will not add much additional new information. In our view it is therefore commendable to build a parsimonious model e.g. by choosing the variable with the highest

change in tau^2 or the one that shows a statistically significant effect in the single risk factor model.

To our knowledge, we have for the first time determined CIs for the estimated change in the between-centre variance tau^2 by successfully applying bootstrapping. In general, research reports just give the percentage of change without any indication of the precision of their estimate. Our results based on 24 diverse study centres, show that the precision is quite low, thus the numbers presented for reduction of tau^2 have to be viewed with a certain caution regarding their absolute values highlighting the explorative nature of this approach. Nevertheless, the magnitude of our standard errors in relation to tau^2 are comparable to those reported by Ivert et al [22] studying the utilization of psychiatric care by 17 729 individuals in 235 neighbourhoods (SE of 0.026 for a tau^2 of 0.052) and Merlo et al [8] investigating survival after heart failure in 4467 patients from 90 hospitals (SE of 0.014 for a tau^2 of 0.053). On the other hand, Gupta et al [21] report a SE of 0.02 for a tau^2 of 0.14 (reported asthma; approx. 45 000 children in 287 neighbourhoods). Thus, overall, large CIs of tau^2 may not be uncommon. Similarly one can see from the literature that the I-squared statistic describing the heterogeneity in meta-analysis often has similarly wide CIs [25]. The imprecision of estimates of tau^2 and its reduction will be especially high if the number of centres is small.

The scaling factors we observe in our data set vary little between models and are closer to one than the ones from the example of simulated data in Bauer [11]. In general, we expect the scaling factor to deviate more strongly from one when the unexplained between-centre variance tau^2 and/or the explained variance $b' \sum X\ b$ are high. The scaling factor will differ between models when the explained variance at individual level differs markedly between the models and/or tau^2 increases: in the case of explained variance at the individual level, an increase in the explained variance will be important because there is no concomitant decrease in tau^2.

Note that when adding centre level variables, no ~~re~~scaling is necessary as these variables cannot explain any of the individual level variation. Thus changes in scaling factors between models will be minute and within the range of imprecision resulting from model estimation.

One limitation of our presented approach is that we have assumed a fixed slope for the individual level variables which is equivalent to assuming the same relationship between influence factors and outcome in all centres. From a methodological point of view, Bauer's method [11] can also accommodate random slopes. The interpretation becomes less straightforward in this case, as the variance of the intercept will depend on the value of the covariate with the random slope and hence the parametrization of that covariate has to be chosen with care. Also, the mere fact of including a covariate will most likely increase the imprecision around the estimated tau^2.

A scaling procedure has also been proposed by Hox [13] based on the approach presented in Snijders & Bosker (2001) [26] which itself is based on McKelvey & Zavoina (1975) [27]: these authors had developed an pseudo R-squared for the latent continuous variable that is very similar to the OLS R-squares. This allows calculating the total variance of the latent variable and is equivalent to the formula given by Bauer [11] for a random intercept model. The scaling factor presented by Hox[13] is obtained by dividing the total variance of the null model by that of the model containing the predictors. Therefore this approach rescales the logistic risk factor model in comparison to the respective logistic (or other non-linear) null model, whereas Bauer's formula rescales the logistic null model and risk factor model to the respective linear underlying latent variable models.

Enzmann & Kohler[14] have developed a Stata ado file named meresc based on the approach presented by Hox[13]. The results for the comparison with the null model obtained by meresc and our Stata-macro are very similar and differ only in the 4th last decimal. This is due to the fact that the procedure to determine the explained variance is slightly different: meresc uses the variance of the predicted values whereas Bauer's formula uses the covariance matrix in combination with the regression coefficients. The equivalence of both approaches to the

explained variance is given in Bauer's article[11] in formula (22). The approach in meresc corresponds to the binary case of the method of Fielding[12] presented for outcomes with several categories. Note that meresc does not run after gllamm.

In conclusion, our approach allows extending the investigation of the influence of personal risk factors on the variation of a continuous outcome between populations to binary outcomes. This is a useful exercise when trying to identify new risk factors especially when they may show little variation in a given location. This method also provides a quantitative insight of the extent to which elimination of a given risk factor would reduce international differences in prevalence. A Stata code is given to implement this kind of analyses.

Tonga); L. García-Marcos (Murcia, Spain); E. von Mutius* (Munich, Germany); U. Keil (Münster, Germany); S.K. Weiland*†, G. Weinmayr* (Ulm, Germany); C.K.W. Lai and G.W.K. Wong (Hong Kong, China); J. Mallol (Santiago, Chile); S. Montefort (Naxxar, Malta); J. Odhiambo† (Nairobi, Kenya); and C. Robertson (Parkville, Australia).

* also members of the ISAAC Phase Two Steering Group, † deceased.

Key messages

Investigating differences in prevalence between centres in a multicentre study, in particular in an international setting, can give important insights regarding relevant risk factors

By quantifying the variation of the random intercept in a multicentre study it can be estimated how much of variation in the outcome prevalence can be attributed to ecological and individual level risk factors

In the case of a binary outcome, e.g. for multilevel logistic regression, a scaling procedure has to be performed to allow the direct comparison between models with different individual level risk factors

---

References:

1. Pearce N. Epidemiology as a population science. *Int J Epidemiol.* 1999;**28**:S1015–S1018.

2. Keys A. Coronary heart disease in seven countries. 41(4 Suppl. Keys A, editor. *Circulation* Circulation; 1970.

3. Asher MI, Stewart AW, Mallol J, et al. Which population level environmental factors are associated with asthma, rhinoconjunctivitis and eczema? Review of the ecological analyses of ISAAC Phase One. *Respir Res.* 2010 Jan;**11**:8.

4. Lee VE, Bryk AS. A Multilevel Model of the Social Distribution of High School Achievement. *Sociol Educ.* 1989;**62**(3):172–192.

5. Klocke A, Clair A, Bradshaw J. International variation in child subjective well-being. *Child Indic Res.* 2014;**7**:1–20.

6. Wainwright NWJ, Surtees PG. Places, people, and their physical and mental functional health. *J Epidemiol Community Health.* 2004;**58**:333–339.

7. Chum A, O'Campo P. Cross-sectional associations between residential environmental exposures and cardiovascular diseases. *BMC Public Health.* 2015 Dec;**15**(1):438.

8. Merlo J, Ostergren PO, Broms K, Bjorck-Linné A, Liedholm H. Survival after initial hospitalisation for heart failure: a multilevel analysis of patients in Swedish acute care hospitals. *J Epidemiol Community Health.* 2001;**55**:323–329.

9. Chaix B, Merlo J, Subramanian S V., Lynch J, Chauvin P. Comparison of a spatial perspective with the multilevel analytical approach in neighborhood studies: The case of mental and behavioral disorders due to psychoactive substance use in Malmö, Sweden, 2001. *Am J Epidemiol.* 2005;**162**:171–182.

10. Snijders TAB, Bosker RJ. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. London etc: Sage Publishers; 2012.

11. Bauer DJ. A Note on Comparing the Estimates of Models for Cluster-Correlated or Longitudinal Data with Binary or Ordinal Outcomes. *Psychometrika.* 2008 Sep 9;**74**(1):97–105.

12. Fielding A. Scaling for Residual Variance Components of Ordered Category Responses in Generalised Linear Mixed Multilevel Models. *Qual Quant.* 2004 Aug;**38**(4):425–433.

13. Hox JJ, Moerbeek M, Schoot R van de. Multilevel Analysis: Techniques and Applications, Second Edition. Taylor & Francis; 2010.

14. Enzmann D, Kohler U. Rescaling Results of Nonlinear Probability Models to Compare Regression Coefficients or Variance Components across Hierarchically Nested Models. *Pap Present Ger Stata Users Gr Meet Berlin.* 2012;

15. Weiland SK, Björkstén B, Brunekreef B, Cookson WOC, Mutius E von, Strachan DP. Phase II of the International Study of Asthma and Allergies in Childhood (ISAAC II): rationale and methods. *Eur Respir J Off J Eur Soc Clin Respir Physiol.* 2004 Sep;**24**(3):406–12.

16. Weinmayr G, Weiland SK, Björkstén B, et al. Atopic sensitization and the international variation of asthma symptom prevalence in children. *Am J Respir Crit Care Med.* 2007 Sep 15;**176**(6):565–74.

17. Strachan DP, Aït-Khaled N, Foliaki S, et al. Siblings, asthma, rhinoconjunctivitis and eczema: a worldwide perspective from the International Study of Asthma and Allergies

in Childhood. *Clin Exp Allergy*. 2015 Jan;**45**(1):126–36.

18.   Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Taylor & Francis; 1994.

19.   Rabe-Hesketh S, Skrondal A. Multilevel modelling of complex survey data. *J. R. Stat. Soc. Ser. A Stat. Soc.* 2006. p. 805–827.

20.   Rabe-Hesketh S, Skrondal A. Multilevel and Longitudinal Modeling Using Stata. Volume II: Categorical Responses, Counts, and Survival. 3rd ed. Stata Press; 2012.

21.   Gupta RS, Zhang X, Sharp LK, Shannon JJ, Weiss KB. The protective effect of community factors on childhood asthma. *J Allergy Clin Immunol*. 2009 Jun;**123**(6):1297–1304.e2.

22.   Ivert AK, Torstensson Levander M, Merlo J. Adolescents' utilisation of psychiatric care, neighbourhoods and neighbourhood socioeconomic deprivation: A multilevel analysis. *PLoS One*. 2013;**8**.

23.   Subramanian S V., Acevedo-Garcia D, Osypuk TL. Racial residential segregation and geographic heterogeneity in black/white disparity in poor self-rated health in the US: A multilevel statistical analysis. *Soc Sci Med*. 2005;**60**:1667–1679.

24.   Vittinghoff E, Glidden D V., Shiboski SC, McCulloch CE. Regression Methods in Biostatistics. Boston, MA: Springer US; 2012.

25.   Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ Br Med J*. 2003;**327**(7414):557–560.

26.   Snijders, Tom A. B.; Bosker RJ. Multilevel analysis : an introduction to basic and advanced multilevel modeling. *Multilevel Anal. an Introd. to basic Adv. multilevel Model.* SAGE, London; 2001.

27.   McKelvey RD, Zavoina W. A statistical model for the analysis of ordinal level dependent variables. *J Math Sociol*. Routledge; 1975 Jan 1;**4**(1):103–120.

Table 1: Single risk factor models: 34809 children in 24 centres

	risk factor prevalence range	OR (CI)	ecological correlation [&]	rescaled tau^2	relative reduction rescaled tau^2 *	tau^2	relative reduction tau^2 *
Empty model (without RF)				0.299		0.329	
Risk factor models							
Mother smoked during first year of life of the child	0.08-43.6%	1.24 (1.14; 1.36)	0.635	0.278	7.63%	0.304	7.17%
Whooping cough infection	0.6-34.2%	1.58 (1.37; 1.83)	-0.097	0.309	-3.95%	0.342	-3.20%
Breastfeeding	26.1-97.8%	0.85 (0.77; 0.94)	0.349	0.306	-2.64%	0.338	-2.27%
Fish consumption	17.2-95.6%	0.89 (0.81; 0.97)	-0.264	0.297	0.74%	0.327	0.77%

* this is the reduction relative to the tau^2 of the empty model: a negative sign shows that there is an increase in tau^2;

§ the prevalence is calculated as $1/(1/\exp(\beta_0)+1)$

& correlation of risk factor prevalence with wheeze prevalence

Table 2: Multivariate risk factor model: 34809 children in 24 centres

		final multivariate model[$]				stepwise
					SE	relative
	risk factor				(rescaled	reduction
Empty model (without RF)	prevalence range	OR (CI)	rescaled tau^2	tau^2)		rescaled tau^2 *
			0.29915	0.099		
Risk factor models						
add Mother smoked during first year of life of the child	0.08-43.6%	1.22 (1.11; 1.33)	0.27769	0.091		7.63%
add Whooping cough infection	0.6-34.2%	1.56 (1.35; 1.80)	0.28722	0.095		3.79%
add Breastfeeding	26.1-97.8%	0.87 (0.78; 0.96)	0.29394	0.098		1.24%
add Fish consumption	17.2-95.6%	0.90 (0.82; 0.98)	0.29214	0.097		1.82%

* this is the reduction relative to the tau^2 of the empty model: a negative sign shows that there is an increase in tau^2;

§ the prevalence is calculated as natural logarithm of the intercept;

$ the odds ratios are given for the final model

Figure legends

Fig 1: Graphical illustration of random intercept model

Fig 2: STATA output for the random intercept model with number of all siblings (num_sibs) as explanatory variable

Fig 3: STATA output for the random intercept model with number of all siblings (num_sibs) and number of older siblings (oldersibs) as explanatory variables

Fig 4. STATA output for the command "correlate num_sibs oldersibs, cov"

Fig 5 a-d: scatterplots of risk factor prevalence vs. wheeze prevalence: a) mother smoking during the first year of life of the child; b) whooping cough infection; c) breastfeeding; d) fish consumption

Figure 1

$$\text{logit}(p_{ij}) = b_{01} + \mathbf{b}\mathbf{X}_{ij}$$

$$\text{logit}(p_{ij}) = b_{02} + \mathbf{b}\mathbf{X}_{ij}$$

$$\text{logit}(p_{ij}) = b_{03} + \mathbf{b}\mathbf{X}_{ij}$$

$$\text{logit}(p_{ij}) = b_{04} + \mathbf{b}\mathbf{X}_{ij}$$

$$\text{logit}(p_{ij}) = b_{05} + \mathbf{b}\mathbf{X}_{ij}$$

One common slope **b** for all centres
One intercept for each centre j:
$$b_{0j} = \gamma_{00} + u_j$$
$$u_j \sim N(0, tau^2)$$

Figure 2

```
Mixed-effects logistic regression              Number of obs    =      27719
Group variable: centre                         Number of groups =         24

                                               Obs per group: min =        289
                                                              avg =     1155.0
                                                              max =       2594

Integration points =    7                      Wald chi2(1)     =      20.40
Log likelihood = -13713.418                    Prob > chi2      =     0.0000

------------------------------------------------------------------------------
       SPT |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
  num_sibs |  -.0542936   .0120221    -4.52   0.000    -.0778565   -.0307306
     _cons |  -1.296155   .1559641    -8.31   0.000    -1.601839   -.9904711
------------------------------------------------------------------------------


------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
centre: Identity             |
                 sd(_cons)   |   .7521476    .112643      .5608227    1.008743
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =  1597.33 Prob>=chibar2 = 0.0000
(obs=27719)
```

Figure 3

```
Mixed-effects logistic regression              Number of obs    =      27719
Group variable: centre                         Number of groups =         24

                                               Obs per group: min =        289
                                                              avg =     1155.0
                                                              max =       2594

Integration points =    7                      Wald chi2(2)     =      20.44
Log likelihood = -13713.388                    Prob > chi2      =     0.0000

------------------------------------------------------------------------------
      sp09x |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   num_sibs |  -.0576607   .0182995    -3.15   0.002    -.0935271   -.0217943
   oldersibs |    .005064   .0207157     0.24   0.807    -.0355379     .045666
      _cons |  -1.295132     .15597    -8.30   0.000    -1.600828   -.9894367
------------------------------------------------------------------------------


------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
centre: Identity             |
               sd(_cons) |    .751893    .1126115     .5606231    1.008419
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =   1594.75 Prob>=chibar2 = 0.0000
(obs=27719)
```
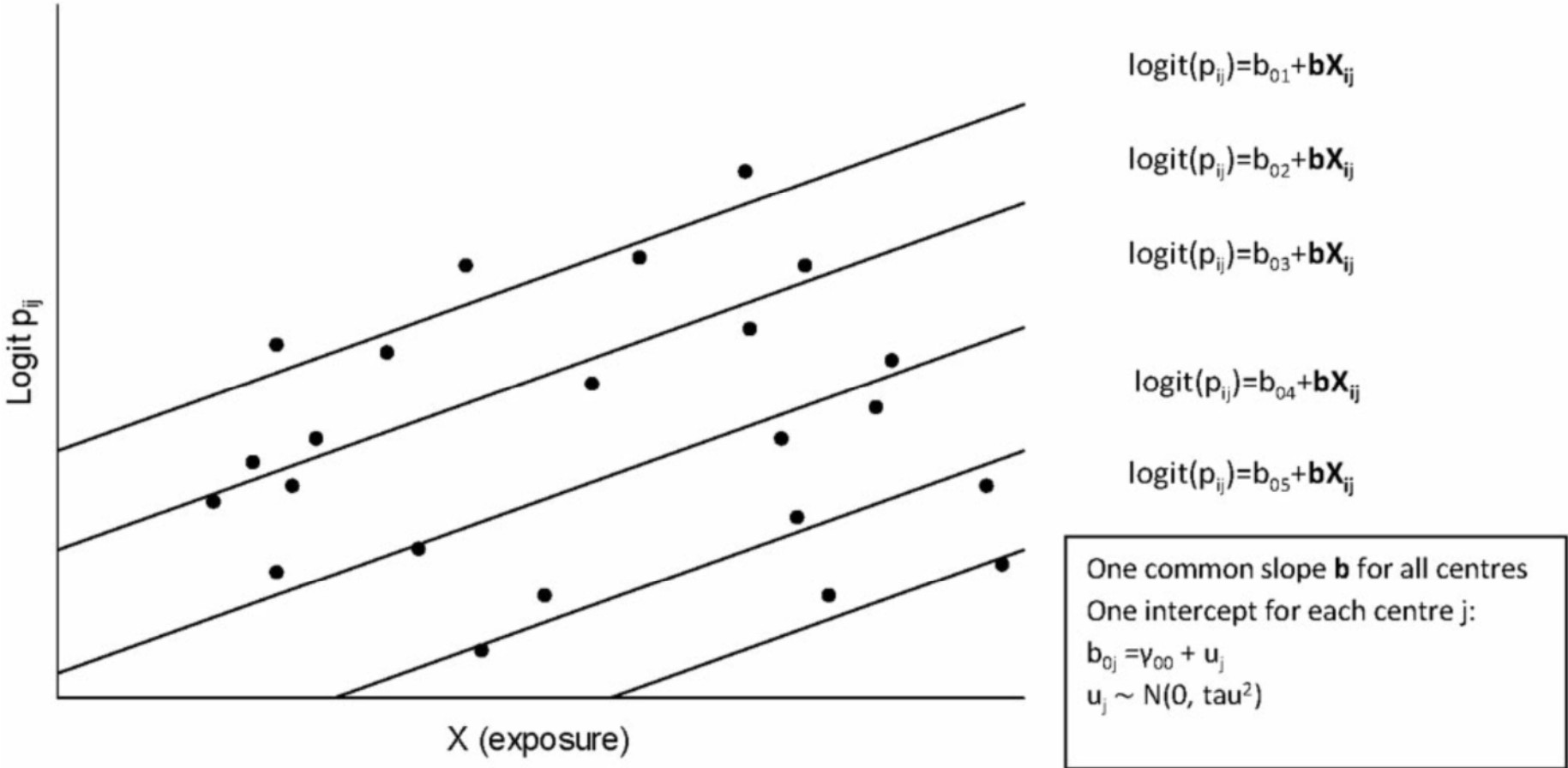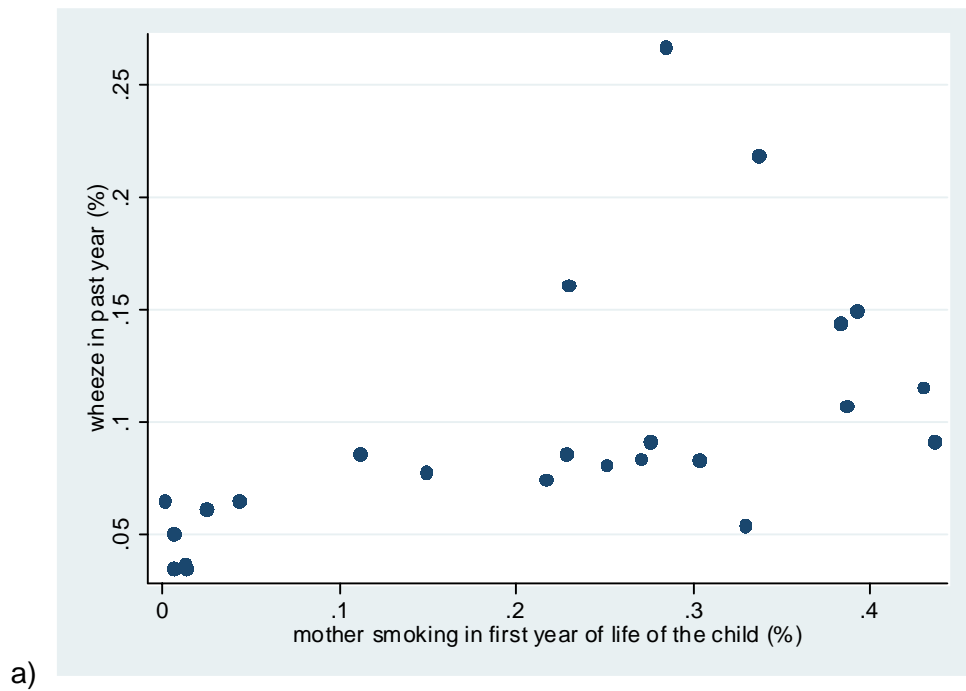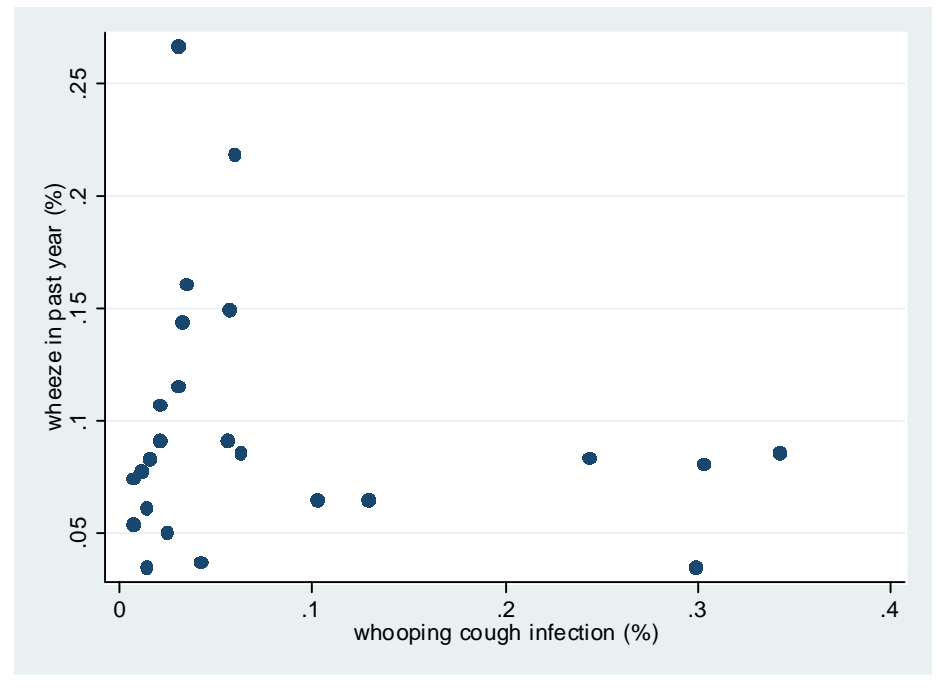
Figure 4

```
             | num_sibs oldersibs
-------------+-------------------
   num_sibs  |  3.19307
  oldersibs  |  1.9859    1.89234
```
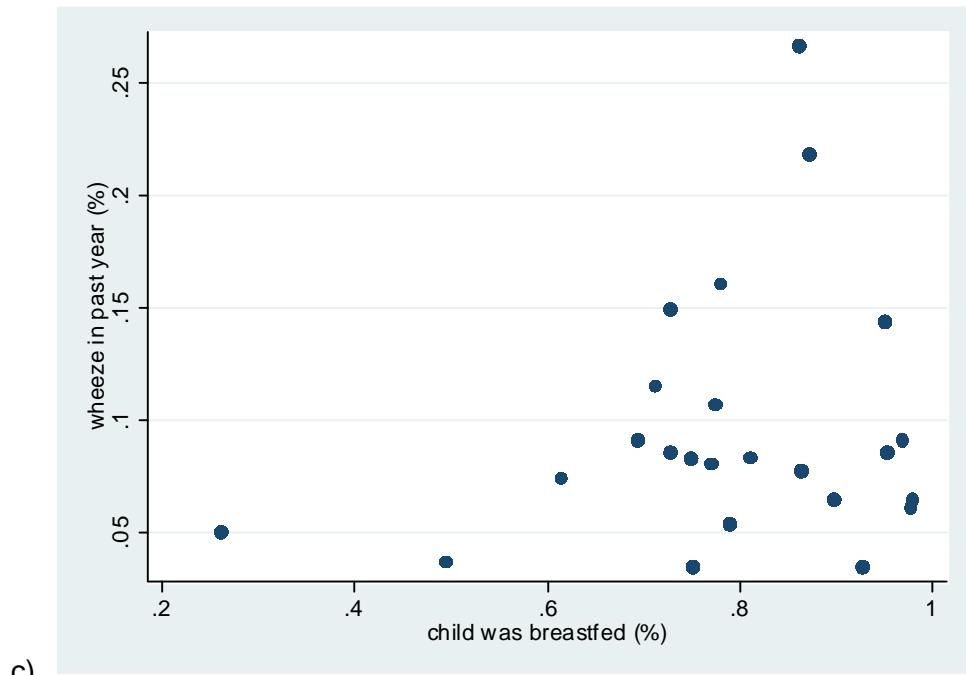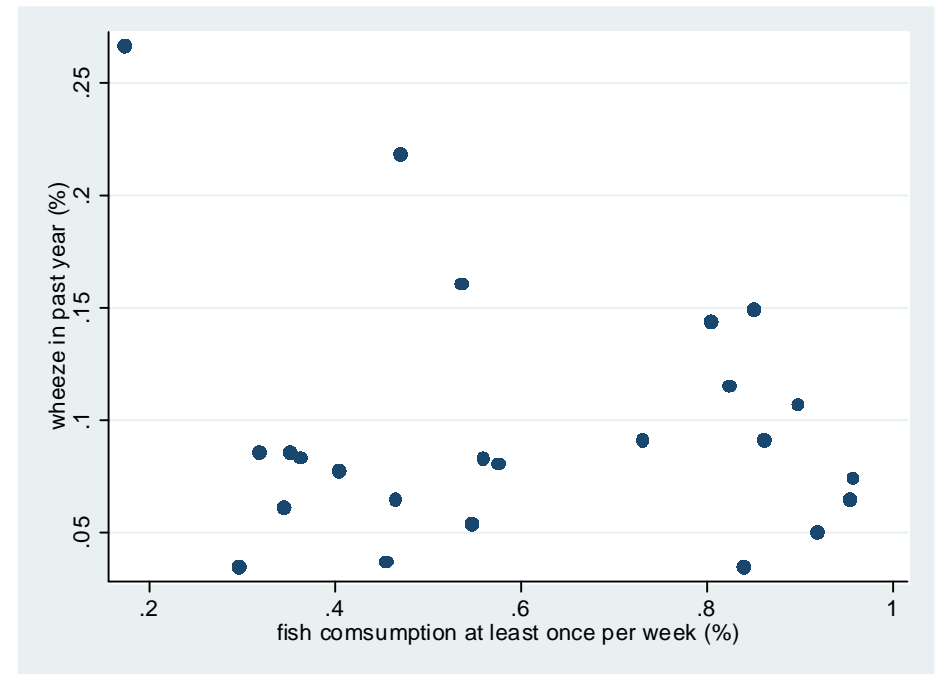
a)

b)

c)

d)

Figure 5

Appendix: Stata macro code

```
* Stata program to rescale estimates obtained from 2-level models (1st
level: individuals 2nd level: study centres)
* calculated in xtlogit
* rescaling formulae from Bauer 2008
* originally programmed in STATA10, later also used in STATA14
**************************************************************************
***********************************************

use "G:\Daten1\ISAACII.dta", clear

cd "G:\Gudrun\ISAACII\Tutorial\"

global logfile="logfilename.log"
log using "$logfile", replace

count

set more off
**************************************************************************
*******************************************
****** regression model for the Null Model (with random intercept for study
centre but without any risk factor)  ***********************
** example: skin prick test positivity = sp09x

xtlogit sp09x, i(centre) quad(30)
matrix u0=e(b)

* tau (NOT tau2) contained in e(sigma_u)
* e(b) contains the individual-level parameters plus lntau (not tau2) as
the last element
* We can't access element of e(b) directly without copying e(b) to another
matrix
matrix u0=e(b)
scalar tau02=e(sigma_u)^2


**** corresponding gllamm syntax**********
** run gllamm twice to speed up estimation, statements nip(5) and then
nip(30) important to get correct estimates!

gllamm sp09xx, i(centre) link(logit) family(bino) nip(5) pweight(wt)
matrix b0= e(b)
gllamm sp09xx, i(centre) link(logit) family(bino) nip(30) pweight(wt) from
(b0)adapt
matrix u0=e(b)
scalar tau02=u0[1,2]^2
****************************************


**** calculate the scaling factor and rescaled tau2 from Bauer, 2008
* calculating null model implied total variance of latent variable y
matrix y0 = tau02 + c(pi)^2/3
mat list y0

* computing the scaling factor(s) to convert to marginally standardized
estimates
scalar scaling_factor_tau02=(c(pi)^2/3)/(y0[1,1])
```

```
scalar rescaled_tau2_null_model = tau02*(c(pi)^2/3)/(y0[1,1])

scalar list


*****************************************************************************
********************************************
***************      regression model with risk factor(s)
*********************************************

* define risk factor sets to be run
global rf_1 "rf06_04xx"
global rf_2 "rf05_04xx"
global rf_3 "rf05_04xx rf06_04xx"
global riskfactor_list " "$rf_1" "$rf_2" "$rf_3" "


* set up temporary file as required by postfile command
tempname testoutfile

foreach rf of global riskfactor_list {

* specify variable types and names you want to have in the output file and
name of the output file
postfile `testoutfile' str100 (riskfactor) float (Nchild Ncentres Tau2
rescTau2 reduction_Tau2 reduction_rescTau2 Relreduction_Tau2
Relreduction_rescTau2 estimate stderr) using "`rf'.dta", replace


****************** regression model with risk factor(s)
*********************************
xtlogit sp09xx `rf', i(centre) quad(30)
matrix u=e(b)
matrix v=e(V)

** extract tau2 and its standard error
scalar tau2=e(sigma_u)^2
scalar list tau2


**** corresponding gllamm syntax**********
** run gllamm twice to speed up estimation, statements nip(5) and then
nip(30) important to get correct estimates!

gllamm sp09xx `rf', i(centre) link(logit) family(bino) nip(5) pweight(wt)
matrix b= e(b)
gllamm sp09xx `rf', i(centre) link(logit) family(bino) nip(30) pweight(wt)
from (b)adapt
matrix u=e(b)
matrix v=e(V)

** extract tau2 and its standard error
scalar tau2=M_cov[1,1]
scalar list tau2
*****************************************


*** Creating covariance matrix required for scaling formula
*****************************
```

```
** = create "cra", the covariance matrix with additional 0-cells needed in
the formula

correlate `rf' if wh02x != ., cov
matrix a=r(C)

*define cn=colsof(a)i.e. number of columns in correlation matrix
scalar cn=colsof(a)
di cn

forvalues i=1/`=cn'{
mat fac = (nullmat(fac)\0)
}
mat list fac

forvalues i=1/`=cn'{
mat c = (nullmat(c)\0)
}
mat list c

mat ca=(c,a)
matlist ca

forvalues i=1/`=cn+1'{
mat r = (nullmat(r),0)
}
mat list r

mat cra=(r\ca)
mat list cra

mat drop fac c r

* Extracting information required for Bauer's scaling formula into a matrix
and its transpose

************ create matrix d  with coefficients  **
** reminder: cn=colsof(a)
mat d = u[1,cn+1]
mat list d
mat list u

forvalues i=1/`=cn'{
*mat c = (nullmat(c)\0)
scalar i2=`i'
scalar list i2
matrix d =  (d\u[1,i2])
}
mat list d


*** calculating model implied total variance of latent variable y
matrix y = d'*cra*d + tau2 + c(pi)^2/3
mat list y


* computing the scaling factor(s) to convert to marginally standardized
estimates
scalar scaling_factor_tau2=(c(pi)^2/3)/(y[1,1])
scalar rescaled_tau2 = tau2*(c(pi)^2/3)/(y[1,1])
scalar list
```

```
*****************************************************************

* extract standard errors of risk factor effect estimate(s)
matrix v=e(V)


post `testoutfile' ("`rf'") (e(N)) (e(N_g)) (tau2) (rescaled_tau2) (tau02-
tau2) (rescaled_tau2_null_model - rescaled_tau2)  ((tau02 - tau2)/tau02)
((rescaled_tau2_null_model - rescaled_tau2)/rescaled_tau2_null_model)
(u[1,1]) (sqrt(v[1,1]))

        ereturn clear

************************
return clear

postclose `testoutfile'

}
**********************************

*** export results from all risk factor models into an excelfile
use "$rf_1", clear
foreach rf of global riskfactor_list {
append using "`rf'"
}

save "sp09x_multivariate"
list
outsheet using "sp09x_multivariate.xls"


log close
**********************************
```