# Investigating Heterogeneity of Effects and Associations Using Interaction Terms

Evangelos Kontopantelis *reader in biostatistics and health services research*1 2, Matthew Sperrin *senior lecturer in health data science*1, Mamas Mamas *professor of cardiology*3 and Iain Buchan *professor in public health informatics*1

1 Centre for Health Informatics, Institute of Population Health, University of Manchester, UK

2 NIHR School for Primary Care Research, UK

3 Keele Cardiovascular group, Institutes for Science & Technology in Medicine and Primary Care, Keele University, UK

Correspondence to:

Evangelos Kontopantelis

Centre for Health Informatics

Institute of Population Health

Vaughan House

Portsmouth Street

University of Manchester

M13 9GB

Email: e.kontopantelis@manchester.ac.uk

## Abstract

Effect heterogeneity, the variability of an association or exposure across subgroups, usually warrants further investigation. The aim of this deeper analysis is to identify effect modifiers (or moderators) and quantify their relationship with the exposure. We explain why it is better to harness interaction effects within a single analytic model than to use separate models to analyse each subgroup. Using examples, we demonstrate a practical approach to modelling and interpretation with interaction terms from various measurement scales (categorical by categorical; categorical by continuous; and continuous by continuous).

## Keywords

interaction terms; effect heterogeneity; effect modifiers; split-sample

## What is new

|  |
| --- |
| * It is often important to investigate the heterogeneity of an effect, i.e. how the effect varies in sub-groups of the population.
* We explain why effect-heterogeneity is best investigated through interaction terms in an analytic model, rather than using separate models.
* Through a series of practical examples we discuss technical aspects and interpretation.
 |

## Introduction

The effect or association of an intervention or exposure might reasonably be expected to vary between subgroups in a population. This heterogeneity of effects can be statistically challenging to investigate,(1) leading to a large number of underpowered investigations and/or selective reporting – often leading to the over-reporting of false positives and false negatives.(2)

Far from being an afterthought, hypotheses about effect heterogeneity might be core to the clinical or epidemiological meaning of a study. In meta-analysis, effect heterogeneity appears to be the norm, and is often of considerable clinical importance.(3) For example, in the quest for “precision medicine” identifying moderator effects with subgroup analyses,(4) i.e. when and for whom an intervention works, should lead to better targeted interventions and more efficient use of resources.

This problem, the need to assess effect heterogeneity but not having the statistical means, is particularly pertinent to randomised clinical trials, which due to large recruitment costs are usually underpowered to detect anything but large effects. Randomised clinical trials are even more underpowered when it comes to subgroup analyses and investigating effect heterogeneity, especially since power calculations are rarely made having subgroup testing in mind.(5) Nevertheless, this shortcoming affects all study designs and highlights a weakness of evidence based medicine in its dependence on P-values.(6) The statistician’s dissatisfaction with significance tests(7) sometimes surprises clinicians until they consider the scope for misinterpretation.(8) Alternative Bayesian approaches can provide more naturally interpreted statistics(9) but they are not a panacea (for example, the choice of prior assumptions may strongly influence the results), so P-values remain a necessary evil for statistical inference, although equivalent information can be much more reliably presented using a confidence or credible interval. What is consistent across all approaches is that, at the very least, the focus should be on the effect size and its clinical relevance. In more recent times, the plethora of routinely collected data has reversed the problem; large observational studies are not underpowered but tend to produce large numbers of statistically significant findings that are not clinically significant, and the solution again is to focus on effect sizes.(10)

Practically, effect heterogeneity can be assessed using either: a single analytic model (e.g. regression) that includes an interaction term between the intervention and the covariate of interest (i.e. effect modifier or moderator); or a split-sample regression. In the latter, a series of models are fit, each model using a subgroup of the data for a specific value of the covariate of interest (e.g. separate models for males and females). From an estimation point of view, the main difference between approaches is the way residual variance is calculated (i.e. standard errors of the estimates) – separately in multiple models, or pooled in interaction models – which in general is relevant only for linear outcomes. The choice between split-sample and interaction approaches usually makes very little difference to inferring the main effects, provided the models do not include additional covariates, or if they do the associations between these covariates and the outcome do not vary for different levels of the effect modifier. However, for detecting effect heterogeneity (i.e. moderator effect) interaction models are more powerful than split-sample regression.(11)

Interaction models have further advantages: First, continuous variables can be used directly, interacting with other continuous variables (e.g. exposure to different radiation levels by age) or with dichotomous variables (e.g. bariatric surgery by BMI level) – avoiding the information loss incurred by categorising the data to enable a separate model approach.(12) Second, researchers often wish to use the model estimates to better quantify and demonstrate effect heterogeneity (e.g. through marginal means), and an interaction model is much more flexible in delivering this through the use of dedicated post-estimation routines, readily available in mainstream statistical software packages. Finally, although statistical significance should be secondary, an interaction model includes a direct estimation of the effect heterogeneity and a comparison of the intervention effect across subgroups, if the effect modifier is categorical. Such comparisons can be manually computed with separate models(13) but they are seldom used and misinterpretations are common (e.g. an effect is statistically significant in model A but not in model B, so the effect is misinterpreted as differing between the models), while differences in estimates may be driven by differences in the covariates of the two groups that are not fully controlled for.

In this paper we explain how to implement and interpret interaction models through practical examples for which we provide relevant Stata code (Online Appendix). We focus on common two-way interactions (i.e. between two variables); although three- or even higher order interactions are possible, they are very difficult to interpret. We interchangeably discuss effect or association heterogeneity, since analytical approaches are identical, but readers should note the difference in interpretation and their relevance to randomised clinical trial or observational studies, respectively.

## Approach

### Categorical by categorical

The simplest categorical contrast is dichotomous by dichotomous, for example to investigate the variation of a yes/no or A/B intervention by sex. The implementation, however, is the same for larger numbers of categories. In a recent analysis we investigated whether the choice of access routes (trans-radial [wrist] or trans-femoral [groin]) for percutaneous coronary intervention (PCI) affected 30-day survival differently across UK health authority areas.(14) This investigation of effect heterogeneity was linked to an important clinical question: “are benefits driven by exceptional cardiologists in certain regions or are clinical gains uniform across all regions”? A random-effects logistic regression model was used to analyse the data, within a multiple imputation setting to account for missing data.(15) The model included both ‘main’ effects of interest, the arterial access route and health authority region, their interaction and many other relevant covariates. Here we only report the main results in Table 1.

The relative benefit of trans-radial over trans-femoral access is reflected in an odds ratio of 0.677 for the North East, the region that is used as a reference category in the analyses (and omitted). The odds ratios for regions then reflect differences in 30-day survival compared to the reference region (North East), but only for the reference access routine (trans-femoral). In other words, the interaction term restricts each of the reported main effects to the reference category in the other variable. Finally, the interaction term shows whether the trans-radial vs trans-femoral effect differs in any other regions compared to the reference region (North East). An odds ratio of 1 for the interaction term would indicate that the effect in specified region was that same as that in the North East. For eight out of ten comparisons, the confidence intervals included 1, indicating no statistically significant differences between these regions and the North East. The non-collapsibility of odds ratios is less of an issue here, where the outcome is relatively rare.(16) However, if the focus is statistical significance, an overall test is first required to assess effect heterogeneity. Such a test was statistically significant (P=0.001) in this case, not surprising in a dataset of 417,038 patients, but judged not to be clinically significant.

If we wished to obtain effect estimates for each region, rather than comparisons to the North East, the point estimates for these effects would be the sum of the coefficient for the North East and the coefficient of the specific difference. For example, the coefficient for the North West would be ln(0.677)+ln(0.933)=-0.45943 and the respective OR=exp(-0.45943)=0.632. However, it is complicated to obtain the standard error of these estimates manually, so we recommend using a post-estimation command to obtain the standard error and produce the confidence intervals. For example, *test* in Stata returns a $χ^{2}$ score, and using this and the coefficient estimate $\hat{β}$, we can compute the standard error as $SE=\hat{β}/\sqrt{χ^{2}}$. Alternative, automated approaches exist for calculating confidence intervals (e.g. *margins* in Stata). Finally, to obtain an overall effect for the intervention the simplest approach is to repeat the analysis without an interaction term.

### Categorical by continuous

Implementing a categorical by continuous interaction is also straight forward, but the interpretation is slightly more difficult to communicate. In a recent meta-analysis of individual-level patient data we investigated whether the initial severity-level of depression influenced the effectiveness of low intensity interventions.(17) In such analyses it is advisable to centre the continuous variable, i.e. its mean subtracted from each observation, in order to ease interpretation. Potential multi-collinearity problems (high correlation between predictors that impedes interpretation) are not alleviated by centring.(18) In our study, we standardised initial depression severity (i.e. centred and divided by the standard deviation so that the final variable had a standard deviation of one) because we had combined numerous measures of depression with varying score ranges and variances. A mixed-effects linear regression model was used to estimate the interaction effect, followed by an individual patient data meta-analysis module to quantify the contribution of each study and display it graphically with a forest plot.(19) Table 2 presents the results summary for the interaction term.

The main effect of -0.42 shows that low intensity treatment is overall more effective than usual care since it leads to a greater reduction on the depression score, on average and for the mean value of the centred initial depression score (i.e. zero). The interaction effect of -0.10 indicates that, on average, a one unit increase in (standardised) initial depression severity is associated with a greater (by 0.10) deterioration in the depression score for low intensity treatment compared to usual care. In other words, the higher the initial depression level of a person, the more effective low intensity treatment is, compared to usual care (Figure 1). The clinical interpretation of this finding is somewhat challenging and it involves a back-transformation (from a standardised score) to a depression scale of choice, to be able to say that an increase in initial severity of one standard deviation on the BDI or CES-D scales would be equivalent to an additional drop of around one point, an effect which may not be clinically significant.(17)

In this scenario, the main effect is reported within the model we used and can be directly interpreted as above. In other words, there is no need to run a separate model to obtain the main effect, unless we wish to remove the interaction term. To obtain simple point estimates of the main effect at different levels of the continuous variable we can simply add the relevant coefficients. For example, to obtain the relative effect of low intensity intervention vs usual care when the standardised depression score is 2 we have -0.42+2\*(-0.10)=-0.62. However, dedicated post-estimation commands can automate the process and easily provide estimates of the effect (and its confidence limits) at various levels of the continuous covariate of interest. Such commands offer great flexibility in calculating adjusted predictions, but care must be exercised with assumptions, for example assuming the data are balanced.

### Continuous by continuous

Most challenging of all to interpret is a continuous by continuous interaction. The Benn index is the coefficient in a regression of ln(weight) on ln(height) and is usually close to 2. Because of this the calculation of BMI as weight/height2 has gained wide acceptance, an approach that ostensibly leads to BMI and height being uncorrelated. In recent work we investigated the Benn Index in more detail, specifically whether it varies depending on gender, age and/or calendar time.(20) In our models, therefore, ln(height) was our primary exposure and we allowed continuous interaction effects between ln(height), age, and calendar time. As well as the complexity of continuous by continuous interactions, we also allowed effects of continuous confounders and interactions to be non-linear, using fractional polynomials,(21) here for clarity we present results for women only from a simpler model with linear effects for centred log(height) and age, and their interaction. The main effect reported for log(height) was 1.58 (95% CI 1.58-1.61) and the interaction effect of log(height) with Y was 0.0035 (95% CI 0.0017-0.0052). Therefore, for each 1-year increase in age the main effect would change by the size of the interaction term. A useful way to illustrate this is a plot of the main effect ln(height) against the value of the effect modifier, age (Figure 2). Alternatively one could plot the predicted relationship between log(height) and log(weight) at different levels of age, in a presentation similar to Figure 1.

## Discussion

We have explained why subgroup analyses using split-regression approaches are not recommended. In addition, we have shown how to investigate effect heterogeneity in a single model with interaction terms, which affords the advantages of: higher power to detect heterogeneity; better covariate control; and greater flexibility in generating estimates and graphs using post-estimation commands. To facilitate model interpretation, analysts need to remember three simple “rules”: 1) include the main effects in the model, not only the interaction term; 2) for binary or categorical variables start coding from zero; and 3) centre interacted continuous variables.

## Sources and selection criteria

We present examples on applications of these methods to a range of research questions and studies as published in major clinical journals, including the BMJ. EK is an experienced biostatistician and health services researcher who, like most statisticians, has used the discussed analytic approaches repeatedly. MS is a statistician and health data scientist with experience in the methods. MM is professor of cardiology who has closely collaborated with EK and MS on projects relevant to the methods. IB is professor of health informatics with wide experience in statistical methodology and its practical implementation.

## Authorship & contributorship

EK wrote the manuscript, with help from MS. MM and IB critically edited the manuscript. EK is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

## Conflicts of interest

MRC Health eResearch Centre Grant MR/K006665/1 supported the time and facilities of EK, MS and IB. No relationships or activities not discussed in the funding statement that could appear to have influenced the submitted work.

## References

1. Altman DG, Matthews JN**.** Statistics notes. Interaction 1: Heterogeneity of effects. BMJ. 1996;313(7055):486.

2. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G**.** Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technol Assess. 2001;5(33):1-56.

3. Kontopantelis E, Springate DA, Reeves D**.** A Re-Analysis of the Cochrane Library Data: The Dangers of Unobserved Heterogeneity in Meta-Analyses. Plos One. 2013;8(7).

4. Wang R, Ware JH**.** Detecting Moderator Effects Using Subgroup Analyses. Prevention Science. 2013;14(2):111-20.

5. Rothwell PM**.** Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet. 2005;365(9454):176-86.

6. Goodman SN**.** Toward evidence-based medical statistics. 1: The P value fallacy. Annals of Internal Medicine. 1999;130(12):995-1004.

7. Sterne JA, Davey Smith G**.** Sifting the evidence-what's wrong with significance tests? BMJ. 2001;322(7280):226-31.

8. Goodman SN**.** A dirty dozen: Twelve P-value misconceptions. Seminars in Hematology. 2008;45(3):135-40.

9. Goodman SN**.** Toward evidence-based medical statistics. 2: The Bayes factor. Annals of Internal Medicine. 1999;130(12):1005-13.

10. Lin MF, Lucas HC, Shmueli G**.** Too Big to Fail: Large Samples and the p-Value Problem. Information Systems Research. 2013;24(4):906-17.

11. Stoneromero EF, Anderson LE**.** Relative Power of Moderated Multiple-Regression and the Comparison of Subgroup Correlation-Coefficients for Detecting Moderating Effects. Journal of Applied Psychology. 1994;79(3):354-9.

12. Altman DG, Royston P**.** Statistics notes - The cost of dichotomising continuous variables. British Medical Journal. 2006;332(7549):1080-.

13. Altman DG, Bland JM**.** Statistics Notes - Interaction revisited: the difference between two estimates. British Medical Journal. 2003;326(7382):219-.

14. Mamas MA, Nolan J, de Belder MA, Zaman A, Kinnaird T, Curzen N, et al.Changes in arterial access site and association with mortality in the United Kingdom: observations from a national percutaneous coronary intervention database. Circulation. 2016.

15. Rubin DB**.** Multiple imputation after 18+ years. Journal of the American Statistical Association. 1996;91(434):473-89.

16. Greenland S**.** Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. Epidemiology. 1996;7(5):498-501.

17. Bower P, Kontopantelis E, Sutton A, Kendrick T, Richards DA, Gilbody S, et al.Influence of initial severity of depression on effectiveness of low intensity interventions: meta-analysis of individual patient data. BMJ. 2013;346:f540.

18. Kromrey JD, Foster-Johnson L**.** Mean centering in moderated multiple regression: Much ado about nothing. Educational and Psychological Measurement. 1998;58(1):42-67.

19. Kontopantelis E, Reeves D**.** A short guide and a forest plot command (ipdforest) for one-stage meta-analysis. Stata Journal. 2013;13(3):574-87.

20. Sperrin M, Marshall AD, Higgins V, Renehan AG, Buchan IE**.** Body mass index relates weight to height differently in women and older adults: serial cross-sectional surveys in England (1992-2011). J Public Health (Oxf). 2015.

21. Sauerbrei W, Royston P**.** Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. Journal of the Royal Statistical Society Series a-Statistics in Society. 1999;162:71-94.

Table 1: Categorical by categorical interaction: arterial access type (artertp2: trans-radial vs trans-femoral) by UK region (geoSHA2) alive/dead status at 30 days (lifestat30d)

| *lifestat30d* | Odds Ratio | Standard Error | p-value | 95% Confidence Interval |
| --- | --- | --- | --- | --- |
| *artertp2* |  |  |  |  |  |
| Radial only | 0.677 | 0.063 | <0.001 | 0.563 | 0.812 |
| *geoSHA2* |  |  |  |  |  |
| North West | 1.048 | 0.108 | 0.651 | 0.857 | 1.281 |
| Yorkshire & Humber | 0.976 | 0.091 | 0.798 | 0.814 | 1.172 |
| East Midlands | 1.112 | 0.109 | 0.275 | 0.919 | 1.347 |
| West Midlands | 1.063 | 0.100 | 0.517 | 0.884 | 1.279 |
| East of England | 0.723 | 0.069 | 0.001 | 0.600 | 0.872 |
| London | 0.961 | 0.083 | 0.645 | 0.811 | 1.139 |
| South East Coast | 0.791 | 0.082 | 0.023 | 0.646 | 0.968 |
| South Central | 0.858 | 0.087 | 0.130 | 0.703 | 1.047 |
| South West | 0.927 | 0.088 | 0.424 | 0.770 | 1.116 |
| Wales | 0.938 | 0.127 | 0.636 | 0.720 | 1.222 |
| *artertp2#geoSHA2* |  |  |  |  |  |
| Radial only#North West | 0.933 | 0.126 | 0.608 | 0.717 | 1.215 |
| Radial only#Yorkshire & Humber | 0.884 | 0.116 | 0.348 | 0.684 | 1.143 |
| Radial only#East Midlands | 0.990 | 0.130 | 0.940 | 0.766 | 1.281 |
| Radial only#West Midlands | 0.960 | 0.118 | 0.737 | 0.754 | 1.222 |
| Radial only#East of England | 1.402 | 0.181 | 0.009 | 1.089 | 1.806 |
| Radial only#London | 1.101 | 0.137 | 0.441 | 0.862 | 1.405 |
| Radial only#South East Coast | 1.465 | 0.229 | 0.015 | 1.078 | 1.991 |
| Radial only#South Central | 1.152 | 0.159 | 0.303 | 0.880 | 1.509 |
| Radial only#South West | 0.898 | 0.111 | 0.385 | 0.705 | 1.144 |
| Radial only#Wales | 0.984 | 0.170 | 0.926 | 0.702 | 1.380 |

Table 2: Categorical by continuous interaction: intervention (group: low intensity vs usual care) by baseline depression score (dept0s) on endpoint depression score (deptFs)

| *deptFs* | Coefficient | Standard Error | p-value | 95% Confidence Interval |
| --- | --- | --- | --- | --- |
| *group* |  |  |  |  |  |
| Low intensity | -0.421 | 0.067 | <0.001 | -0.553 | -0.288 |
| *group#c.dept0s* |  |  |  |  |  |
| Low intensity | -0.097 | 0.048 | 0.045 | -0.192 | -0.002 |

Figure 1: Categorical by continuous interaction: intervention (low intensity vs usual care) by baseline depression score on endpoint depression score\*

\*post-estimation predictions based on a slightly different model to the one reported: original meta-analysis of individual patient data modelled baseline depression as study specific and fixed, calculating different effects for each study (here we use a model that calculates one estimate across all studies)

Figure 2: Illustration of continuous by continuous interaction. Age (the effect modifier) varies in the x-axis while the y-axis shows the log(height) coefficient at that age, and its 95% confidence intervals.

