Rebecca Hein
Dr. sc. hum.

Statistical Methods for the Investigation of Gene-Environment Interactions in Genetic Epidemiological Association Studies

Geboren am 28.02.1981 in Trier
Diplom der Fachrichtung Wirtschaftsmathematik am 21.09.2005 an der Technischen Universität Kaiserslautern

Promotionsfach: DKFZ (Deutsches Krebsforschungszentrum)
Doktormutter: Frau Prof. Dr. sc. hum. Jenny Chang-Claude

This thesis focuses on the investigation of gene-environment (GxE) interactions in case-control studies of candidate genes. When true disease variant(s) are not genotyped, these variant(s) are searched for by indirect association using genetic markers in linkage disequilibrium (LD) with the true disease variant(s). Single markers capture information only on pairwise LD, whereas haplotypes comprise information on pairwise as well as on multi-marker LD. Thus, regions may exist where single-marker approaches fail to detect associations to the disease, whereas it may be possible to detect these associations using haplotype-based methods. The aims of this thesis were threefold:

1) Power and sample size estimation for indirect association studies of gene-environment interactions: An algorithm for power and sample size estimations in indirect matched case-control studies of GxE interactions was derived and implemented (http://www.dkfz.de/en/klepidemiologie/software/software.html). Using this algorithm, the influence of the following variables on sample sizes required to detect GxE interactions were explored: 1) true genetic and environmental main effects, 2) disease allele frequencies, 3) matching of marker and disease allele frequencies, 4) the LD between the markers and disease variants, 5) prevalence of environmental exposures, and 6) the magnitude of interactions. Additionally, sample sizes required to detect genetic marginal and GxE interaction effects in direct and indirect association studies were compared.
Required sample sizes may be heavily inflated if LD between marker and disease variants decreases. More than 10,000 case-control pairs may be required to detect GxE interactions. However, given weak true genetic main effects, moderate prevalence of environmental exposures, as well as strong interactions, GxE interaction effects may be detected with smaller sample sizes than those needed for the detection of genetic marginal effects. Moreover, in this scenario, rare disease variants may only be detectable when GxE is included in the analyses.

2) Comparison of the power of single-marker versus haplotype-based association methods for gene-environment interactions: The power of single-marker versus haplotype-based methods for analyzing GxE interactions was compared using simulated and real data on rheumatoid arthritis.
In the analysis of the simulated data, stepwise and regular conditional logistic regression (CLR) was performed using a matched case-control sample. Haplotype-based analyses were performed using the Mantel statistics using haplotype sharing and a prospective haplotype-trait association test. The real data were investigated using logistic regression (LR)/CLR and the two haplotype-based methods described above, using 1) a sample of about 1,000 unrelated individuals and 2) a small family-based sample.

For markers in strong LD, stepwise CLR performed poorly because of the correlation between the predictors in the model. The power was high for detecting genetic marginal effects using simple CLR models and haplotype-based methods and for detecting joint effects using CLR and Mantel statistics. When main and interaction effects were included into the model, due to the strong main effect of the risk variant, the GxE interaction effect was not detectable using CLR. By contrast, the prospective haplotype-trait association test had moderate to high power to detect the GxE interaction. In the real data sample of unrelated individuals investigating a genetic region characterized by strong LD, a previously reported risk variant (R620W) in the PTPN22 gene was identified using LR and the prospective haplotype-trait association test. Moreover, the previously reported R620W-sex interaction was confirmed by LR analysis. In the small family sample (low LD between markers), using CLR, the genetic marginal effect of a previously identified variant (DRB1) in the HLA region could be confirmed.

The results suggest that stepwise or other automated variable selection methods are not suitable for the investigation of GxE interactions in regions with high LD, as measured by D'. Given strong genetic marginal effects, moderate GxE interactions, and moderate sample sizes, as in the simulated data, single-marker and haplotype-based methods all had adequate and comparable power to detect genetic marginal effects without considering the GxE interaction. The findings regarding joint effects suggest that analysis of joint effects may be an option for detecting genetic effects in some situations. In the simulated scenario, the prospective haplotype-trait association test had better power than simple CLR modeling to detect moderate GxE interactions.

3) Comparison of the type I error and the power of different haplotype-based association methods for gene-environment interactions: The investigated four methods employ different versions of the expectation maximization algorithm to estimate haplotype frequencies and differ in the choice of the reference group as well as in the way the risk of disease is modeled (retrospective versus prospective and joint likelihood approaches). Assumptions such as Hardy Weinberg equilibrium (HWE), a specific mode of inheritance, gene-environment independence, and a rare disease are imposed by the different methods.

Direct and indirect association scenarios were generated, where haplotype pairs were either simulated to be in HWE or Hardy Weinberg disequilibrium (HWD). The two retrospective methods were more powerful for detecting GxE interactions than the other investigated methods. However, under HWD, in contrast to the prospective and joint likelihood method, the retrospective methods, did not strictly abide by the nominal significance level. The power in the indirect scenarios was generally decreased compared to the direct scenarios. However, since haplotypes capture information on pairwise as well as on multi-marker LD, the power of the haplotype-based methods may still be adequately high in the indirect scenarios. The superior performance of the retrospective methods with respect to power may be ascribed to different factors. Firstly, these methods account for the sampling design of the case-control data. Secondly, the efficiency of these methods is increased, since they exploit the rare disease and the gene-environment independence assumption.

Accounting for GxE interactions in genetic association studies may improve the power to detect genetic effects and may help to identify important environmental effect modifiers. For the investigation of GxE interactions in case-control data, retrospective haplotype-based methods may be an attractive alternative to haplotype-based methods, which do not account for the case-control ascertainment, as well as to single-marker approaches.