# Rare disease and other assumptions in statistical analysis of genetic data: the good, the bad, and the ugly

Shili Lin*
The Ohio State University, Columbus, OH, USA - shili@stat.osu.edu

Jingyuan Yang
The Ohio State University, Columbus, OH, USA - summerice0510@gmail.com

Statistical applications in genetics, pioneered by R. A. Fisher, have been instrumental in the discoveries of genes and genetic variants predisposed in genetic diseases. To facilitate such applications, various assumptions are frequently made, although their effects are sometimes not well investigated, or worse, the assumptions are "forgotten" and therefore not checked before the application of a statistical procedure. Population genotype frequency distribution, such as Hardy-Weinberg equilibrium (HWE), is often assumed in many statistical procedures, so is the rare disease assumption. One particular problem where such assumptions are made is in the investigation of maternal (genotype) effect in the context of genome-wide association study (GWAS). Maternal effect refers to the phenomenon that the mother's genotype is being expressed in her child's phenotype, regardless of whether the mother's gene was actually passed to the child, leading to an important type of parent-of-origin expression patterns. An effective design for detecting maternal effect is that of case-mother/control-mother, as such a design eliminates the need to genotype sometimes hard-to-recruit fathers. A log-linear model has been proposed to analyze the data to assess the existence of maternal effect, but a number of assumptions are made to avoid overparameterization. They include mating symmetry and allelic exchangeability (weaker conditions than HWE), and the disease being rare. In this paper, we investigate the effects of these assumptions on power, type I error, and biases of parameter estimates based on the log-linear model. We show that *good* results, with an increase in power, may indeed be obtainable under certain situations. However, the results can be *bad*, leading to a substantial amount of bias even if the assumptions are met. When some of the assumptions are violated, the results are *ugly*; there can be severely inflated type I errors and huge biases. In contrast, we offer a logistic model as an alternative approach for detecting maternal effects for data from the same design. Our simulation study show that the type I error rates are well controlled without compromising much power, regardless of whether the assumptions hold or not.

**Keywords**: Bias and type I error, maternal effect, rare disease assumption, Hardy-Weinberg equilibrium.