

# Confounding by Population Stratification A Guided Reading in Genetic Epidemiology

## Educational Objectives

After reading this guided reading, you should be able to:

- Describe what is meant by population stratification.
- Understand the conditions under which population stratification can occur.
- Discuss how population stratification may affect the interpretation of case-control genetic association studies.
- Be familiar with the methods used to address population stratification.

## Population Stratification Overview

This learning module is aimed primarily at epidemiologists, but could also be adapted for discussion among public health professionals.

**C**ase-control association studies are a widely used study design in genetic epidemiology. Under this method allele and/or genotype frequencies at a genetic marker are compared between cases and controls. Association studies are thought to be well suited for identifying genetic variation underlying complex traits, such as diabetes, cancer and cardiovascular disease. However, this study design is susceptible to potential confounding by population stratification.

Population stratification stems from the fact that populations are typically heterogeneous in terms of genetic ancestry. For example, a population may comprise two or more groups with distinct genetic ancestry, such as different ethnic groups living in a U.S. city. Alternatively, there may have been genetic mixing of two or more

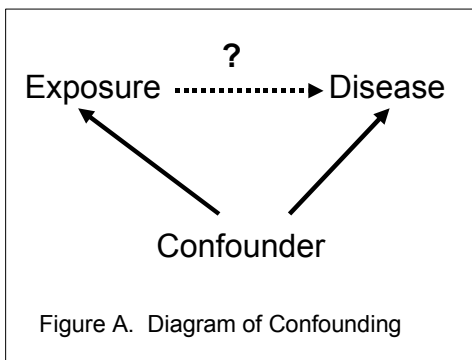


Figure A. Diagram of Confounding

groups in the recent past. This later situation, termed genetic admixture, is seen in Native American populations, who have both American-Indian and Caucasian ancestry. Allele frequencies are known to vary among populations of different genetic ancestry, and, similarly, disease risk often varies among populations of different genetic ancestry. This is akin to the situation of confounding. A confounder is defined as a factor that is associated with both the exposure and the disease, but is not a result of the exposure (Figure A). Confounding factors may bias the observed association between an exposure and disease. In this same manner, differences in allele frequencies between cases and controls may, in fact, be spurious associations resulting from differences in genetic ancestry.

Using a classic study by Knowler et al, we will detail how population stratification can confound the relationship between a genetic marker and disease. We will discuss how population stratification can influence the results and interpretation of a genetic association study. Finally, we will list recent review papers that address this issue in more detail.

This guided reading assumes that the reader has read the article by W. C. Knowler, R. C. Williams, D. J. Pettitt and A. G. Steinberg. **Gm<sup>3;5,13,14</sup> and Type 2 Diabetes Mellitus: An Association in American Indians with Genetic Admixture.** Am J. Hum. Genet. 43; 520-526, 1988.

## Summary of Article

**K**nowler et al. (1988), collected data on 4,920 Pima and Papago Indians as part of a large longitudinal study. The study began in 1965, and was motivated by the high prevalence of diabetes in these Native American populations. For each participant the researchers determined type II diabetes status (using the World Health Organization criteria), self-report of fraction of American Indian heritage, and Gm genotypes.

The Gm system of human immunoglobulin G (IgG) is a highly variable marker, often used to study gene flow among human populations. Polymorphisms in this system are characterized by variants in the genes coding for the heavy chains of IgG. This type of marker was commonly used for genetic association studies in the 1980's. Today, genetic association studies often use single-nucleotide polymorphisms (SNPs) as markers. For the purpose of this exercise, the "Gm<sup>3;5,13,14</sup> haplotype" may be treated like a biallelic marker, similar to a SNP. Individuals either have this particular Gm variation or they do not. For this guided reading, we will refer to this variant as the Gm marker.

Knowler et al. observed a negative association between the Gm marker and type II diabetes. However, further analyses in the paper show that the association results from confounding by genetic admixture. Specifically, the Gm marker is more common and the prevalence of type II diabetes is lower in individuals with no American Indian ancestry compared to individuals of full American Indian ancestry. When Knowler et al. stratified on level of American Indian ancestry, they no longer saw an association between the Gm marker and type II diabetes. This is the best demonstration to date of the potential for false associations due to population stratification. As a result, this paper has had a lasting legacy in the field of genetic epidemiology.

## Questions on the Article

The authors of this paper present prevalence ratios as a measure of excess risk. In the traditional epidemiological 2x2 table (see Figure B) the prevalence ratio is:

		Disease	
		+	-
Exposure	+	a	b
	-	c	d

Figure B. Generic Two-by-two table

$$\left(\frac{a}{a+b}\right) / \left(\frac{c}{c+d}\right).$$

Most current genetic association studies present results as an odds ratio, and the association between Gm marker and type II diabetes can also be summarized with that measure. The formula for an odds ratio is:

$$\left(\frac{a}{b}\right) / \left(\frac{c}{d}\right) = (a * d) / (c * b).$$

		Diabetes	
		+	-
Gm Marker	+	23	270
	-	1343	3284

Figure C. Two-by-two table for Table 2 of Knowler

The information presented in Table 2 of the Knowler paper can be rewritten in the form of a 2 x 2 table (see Figure C). Using this form, it is straightforward to calculate the prevalence ratio:

$$\left(\frac{23}{23+270}\right) / \left(\frac{1343}{1343+3284}\right) = 0.27.$$

And to calculate the odds ratio:

$$\left(\frac{23}{270}\right) / \left(\frac{1343}{3284}\right) = 0.21.$$

As with any epidemiological study, we conclude that we have evidence for an association between a genetic marker and disease if the odds ratio or prevalence ratio is significantly different from 1.

**Question 1:** Using only the information in Table 2 of the Knowler paper, what might you conclude about the GM marker and risk of diabetes?

**YOUR ANSWER:**

**Question 2:** Using only the information in Table 4 of the Knowler paper, quantify the association between the Gm marker and type II diabetes. What might you conclude from this result about the GM marker and risk of type II diabetes?

**YOUR ANSWER:**

The results and interpretation of the association between the GM marker and diabetes differ for the entire population (Table 2, Knowler paper), as compared to the age and heritage restricted population (Table 4, Knowler paper). The observed association for the entire population may result from confounding by genetic admixture.

**Question 3:** What is confounding?

**YOUR ANSWER:**

**Question 4:**

Looking at Figure 3 of the Knowler article, what does the left hand panel indicate about the association between level of American Indian heritage and outcome (type II diabetes)? What does the right hand panel indicate about the association between level of American Indian heritage and exposure (Gm haplotype)? Are the conditions for confounding by heritage met in this situation? Why or why not?

**YOUR ANSWER:**

Table 3 of the Knowler paper presents the prevalence of type II diabetes stratified both by age and by American Indian heritage. The last row gives the age-adjusted prevalence, both with and without the Gm marker, for different levels of American Indian heritage. The adjusted prevalence is a weighted average of the prevalence in each of the different age categories.

Table A: Age-adjusted Prevalence of type II Diabetes according to Gm Marker, stratified by level of Indian Heritage (taken from the last two rows in Table 3 of Knowler et al).

<i>Gm marker</i>	<i>Indian Heritage (in Eighths)</i>		
	<i>0</i>	<i>4</i>	<i>8</i>
Present	17.8	28.3	35.9
Absent	19.9	28.8	39.3

**Question 5:** Using this information, what is the age-adjusted prevalence ratio for individuals of full American Indian heritage? 50% American Indian heritage? Full Caucasian heritage? What would you expect to be the approximate prevalence ratio if you adjusted for both age and American Indian heritage (you do not need to do actual calculations, just give an approximate answer)?

**YOUR ANSWER:**

**Question 6:** Overall, how do you interpret the findings of this study? What are the implications of the findings (if any) for genetic association studies? What are the implications of the findings (if any) for public health?

**YOUR ANSWER:**

# Conclusions

**P**opulation Stratification (in this case due to admixture of two populations) can lead to spurious associations between a genetic marker and a disease phenotype. In this study, a crude comparison showed an inverse association between the Gm marker and type II diabetes. However, a stratified analysis comparing individuals with the same degree of American Indian heritage showed no association between the Gm marker and type II diabetes. In this study degree of admixture, as measured by fraction of American Indian ancestry, was a confounder. Specifically, level of American Indian ancestry was associated with both the Gm haplotype and type II diabetes. Failure to properly control for this confounding variable resulted in a spurious association. If the researchers hadn't considered degree of American Indian heritage in their analysis, they would have reached a false conclusion about the association between the Gm marker and type II diabetes

## How Important is Population Stratification?

In the 1990's concern over population stratification lead some researchers to question the validity of population based genetic association studies. Currently the impact of population stratification is up for debate. Recent review papers have argued both for (Ziv and Buchard 2003, Thomas and Witte 2002) and against (Cardon and Palmer 2003, Wacholder, Rothman and Caporaso 2002) population stratification as a major source of bias.

The general consensus of these papers is that bias due to population stratification will likely be small in well designed and analyzed case-control studies of Caucasian populations. Several methods exist for controlling for population stratification. Researchers can use traditional epidemiological methods; such as matching, restricting or adjusting for ethnic background. They can also use newer methods specific to genetic epidemiology; such as a) family-based methods including transmission disequilibrium tests (TDT), or b) statistical methods using unlinked genetic markers to detect, quantify and correct for stratification. Each of these methods has strengths and limitations.

Population stratification, and the methods used to control for it, should be addressed in any population-based genetic case-control study. Interested parties should consult the review articles and Med-Line Search opportunity to learn more about the theoretical and empirical research being done in this area.

# Bibliography

### *Guided Reading Paper:*

Knowler W.C. Williams, R.C. Petitt, D.J. and Steinberg, A.G. Gm<sup>3;5,13,14</sup> and Type 2 Diabetes Mellitus: An Association in American Indians with Genetic Admixture. Am. J. Hum. Genet. 43; 520-526, 1988.

### *Review Articles:*

Cardon, L. R. and Palmer, L. J. Population Stratification and Spurious Allelic Association. The Lancet. 361; 598-604. 2003.

Thomas, D. C., and Witte, J. S. Point: Population Stratification: A Problem for Case-Control Studies of Candidate-Gene Associations? Cancer Epidemiol. Biomark. Prev. 11; 505-512. 2002.

Wacholder, S. Rothman, N and Caporaso, N. Counterpoint: Bias from Population Stratification is Not a Major Threat to the Validity of Conclusions from Epidemiological Studies of Common Polymorphisms and Cancer. Cancer Epidemiol. Biomark. Prev. 11; 513-520. 2002.

Ziv, E. and Burchard, E. G. Human Population Structure and Genetic Association Studies. Pharmacogenomics. 4; 431-441. 2003.