Collection of Biostatistics Research Archive

COBRA Preprint Series

Year 2012 *Paper* 94

Why odds ratio estimates of GWAS are almost always close to 1.0

Yutaka Yasui*

*University of Alberta, yyasui@ualberta.ca

Why odds ratio estimates of GWAS are almost always close to 1.0

Yutaka Yasui

Abstract

"Missing heritability" in genome-wide association studies (GWAS) refers to the seeming inability for GWAS data to capture the great majority of genetic causes of a disease in comparison to the known degree of heritability for the disease, in spite of GWAS' genome-wide measures of genetic variations. This paper presents a simple mathematical explanation for this phenomenon, assuming that the heritability information exists in GWAS data. Specifically, it focuses on the fact that the great majority of association measures (in the form of odds ratios) from GWAS are consistently close to the value that indicates no association, explains why this occurs, and deduces two specific forms of epistasis/interaction as its cause. The implication is that GWAS may be able to recover "missing heritability" if the two specific forms of epistasis and gene-environmental interaction are fully explored.

1. INTRODUCTION

A genome-wide association study (GWAS) of disease susceptibility compares "cases" who have developed a disease of interest to "controls" who have not, within a defined study population, with respect to genetic variations. Most commonly, the genetic variations examined in GWAS are those of single nucleotide polymorphisms (SNPs) across the whole genome. As such, a GWAS is an epidemiological case-control study with SNP genotypes being the "exposures" of interest that may modify the disease risk. The measure of association typically used in case-control studies is odds ratio (OR) [Gordis, 2008]. While the underlying odds of developing the disease in a defined time period (i.e., the case-control ratio) is fixed by the case-control study design, and thus not estimable, its ratio between two groups with and without a hypothesized cause of the disease is estimable [Cornfield, 1951]. Specifically in GWAS, ORs of developing the disease are estimated between individuals with and without certain genotypes of SNPs: departures of the ORs from 1.0 indicate that corresponding genotypes of SNPs (or the regions of genome they represent) are associated with the disease risk.

GWAS has become a major research framework, often conducted as multi-country collaborative projects with great resource requirements, aiming to identify regions of human genome whose variations across subjects are associated with risk of various diseases. This aim has been accomplished only partially, however, even for diseases that have been investigated by a large number of GWASs such as Crohn's Disease. Specifically, the known degree of genetic contributions to the risk of developing a particular disease (i.e., the "heritability"), known from other types of studies (e.g., family studies), is explained poorly by GWAS findings, where only a small fraction of the known degree of the "heritability" are attributable to SNP variations discovered by GWASs. This problem is referred to as the "missing heritability" problem and its potential causes have been debated widely [Manolio *et al.*, 2009; Goldstein, 2009].

The aim of this paper is to attempt to explain the "missing heritability" problem with precise statistical reasoning and provide a potential direction with which the problem can be tackled. The strategy is to apply a key statistical theorem on ORs in the population, not in the sample, namely, the collapsibility theorem for the analysis of contingency tables, to show that the "missing heritability" problem is due to two specific forms of interaction that are not widely assessed in GWAS analysis. The implication of this work is that, assuming GWAS data contain the "heritability" information, its extraction must target these specific forms of interaction.

2. COLLAPSIBILITY AND MARGINAL ODDS RATIOS

Analysis of GWAS data is typically performed for a single SNP, one at a time, without involving any other SNPs or non-genetic factors. The salient systematic feature of GWAS findings, regardless of diseases and populations studied, is that the great majority of resulting OR estimates are close to 1.0. These near-null OR estimates lead to, and is the essence of, the "missing heritability" problem of GWAS. This characteristic of GWAS OR estimates cannot be attributable to stochastic properties of the OR estimation, including statistical power: if it could, OR estimates would be highly variable and not consistently close to 1.0. Thus, underlying ORs in the population must also be close to 1.0. If the underlying biology in the population were such that there is only one causal factor for the disease and it is the genotype of a single SNP (or the region of genome it represents), the marginal OR of the disease defined for the SNP genotype in the population, ignoring all the other SNPs and non-genetic factors, would represent the underlying biological disease-SNP association accurately. The underlying biology of complex diseases studied in GWAS, however, is expected to involve multiple SNPs and non-genetic factors. With the existence of multiple causal factors, the marginal OR of the disease defined for one of the causal factors would not represent accurately the extent and mechanism of the underlying disease biology involving the factor.

Statisticians have proven theorems that specify exact circumstances under which one can ignore other relevant factors and assess the marginal association between two variables accurately (when all variables are categorical) [Simpson, 1951; Whittemore, 1978; Ducharme and LePage, 1986]. Using relevant terms in GWAS, one of such theorems can be stated as follows.

The SNP-disease association measured by the marginal OR between genotypes of a single SNP and the case-control status of a disease, without considering any other variable, is equal to the association measured by the conditional OR between these two variables in any subgroup defined by a third variable (and therefore the third variable can be ignored in the assessment of the SNP-disease association) if and only if the SNP genotype is independent of the third variable in both cases and controls, or the third variable is independent of the case-control status in all SNP-genotype groups.

Note that “the third variable” can be a combination of multiple variables including SNPs and non-genetic factors. According to the theorem, the marginal OR calculated for the genotype of a single SNP, ignoring all the other relevant factors, genetic and non-genetic, is accurately representing the SNP-disease association in any subgroups defined by the other relevant factors, if and only if all the other relevant factors are independent of the SNP genotype in both cases and controls (or the trivial case holds where there is no other factor other than the SNP associated with the disease). This theorem can be proven for the GWAS scenario, using elementary algebra of OR as shown in Appendix A. Without loss of generality, we will consider hereafter binary genotypes of a SNP (corresponding to dominant or recessive inheritance).

Given that even the largest OR estimates are close to 1.0 in the great majority of GWASs of disease susceptibility and they do not explain the full extent of known degrees of disease heritability, the theorem suggests that the violation of its condition is the norm in the underlying biology of the complex diseases studied in GWAS. How is the theorem’s condition violated? In a properly-designed case-control GWAS, Mendelian Randomization [Clayton and McKeigue, 2001] implies that genotypes of a SNP are independent of “the third variable” in controls, unless “the third variable” is genetic and in the region of linkage disequilibrium (being correlated) with the SNP. Thus, the theorem’s condition that is violated has to be the independence of the SNP genotype and “the third variable” in cases. In other words, the theorem identifies a potential source of the “missing heritability” problem to be the dependence of the SNP genotype and “the third variable” in cases.

3. PRECISE CONDITIONS THAT MAKE ODDS RATIOS CLOSE TO 1.0

Let us quantify this dependence in cases and identify when the marginal OR estimate of a SNP gets close to 1.0 in spite of its strong association (not as a single SNP, but together with “the third variable”) with the disease. To simplify the discussion without loss of generality, consider a binary “third variable” and indices i and j for the i^{th} genotype of the SNP of interest (i = 0, 1) and the j^{th} category of “the third variable” (j = 0, 1).

The independence of the SNP genotype and “the third variable” in cases implies that there is no interaction between the two, i.e., the conditional OR of the disease comparing the SNP genotype 1 vs. 0, conditioned on “the third variable”, j = 0 or 1, does not change between the categories of “the third variable”, i.e.,

$OR_{1|j} = OR_{1|0}$, where $OR_{1|j}$ is the conditional OR of the disease comparing the SNP genotype 1 vs. 0, conditioned on being in the *j*th category of "the third variable". Thus, the departure from the independence can be expressed by $OR_{1|1} = OR_{1|0}(1+\delta)$ with non-zero δ (>-1). Then, as shown in Appendix B, the marginal OR, denoted by *OR*, that is estimated in the single-SNP GWAS analysis ignoring "the third variable" is given by:

$$OR = OR_{1|0} \times \left\{ 1 + \delta \Big/ \left(1 + \frac{p_{00} \times odds_{00}}{p_{01} \times odds_{01}} \right) \right\},$$

where p_{ij} be the proportion of subjects with the *i*th genotype of the SNP and the *j*th category of "the third variable" among controls, and $odds_{ij}$ be the ratio of this proportion among cases to p_{ij}, the same proportion among controls.

According to the equation, the marginal OR is close to 1.0 systematically in the population, in spite of the SNP's association with the disease risk in only two scenarios: (A) $OR_{1|0}$ is greater than 1.0 but δ is negative and thus the marginal OR gets attenuated towards 1.0; and (B) while $OR_{1|0}$ is close to 1.0, $OR_{1|1}$ is not with a positive δ, but the factor in the parenthesis ( ) of the equation is large and thus the marginal OR gets close to 1.0. These correspond to the following two specific forms of interaction.

Interaction of Redundancy/Masking

This form occurs when the SNP is strongly associated with the increased disease risk in the absence of "the third variable" ($OR_{1|0}$ is greater than 1.0), but the presence of "the third variable" masks the SNP-disease association. The marginal OR gets close to 1.0 under this scenario if "the third variable" is prevalent (in the SNP=0 group) and/or strongly positively associated with the disease (in the SNP=0 group).

For example, under so-called "genetic heterogeneity", two or more SNPs may be associated with disease risk such that either is sufficient to modify disease risk, but neither is necessary. They are redundant and the presence of both does not result in the sum of the two effects. These are illustrated by broken lines in

Figure 1. As a specific numerical example, $OR_{1|0} = 10.0 = odds_{01} / odds_{00}$, $\delta = -0.9$, and $p_{01} = \gamma \times p_{00}$ would yield the marginal OR = 1.8, 1.2, and 1.1 for γ = 1, 5, and 9, respectively. The OR of 10.0 for each SNP in the absence of the other can be attenuated to the marginal OR that is close to 1.0, depending on the prevalence of the other.

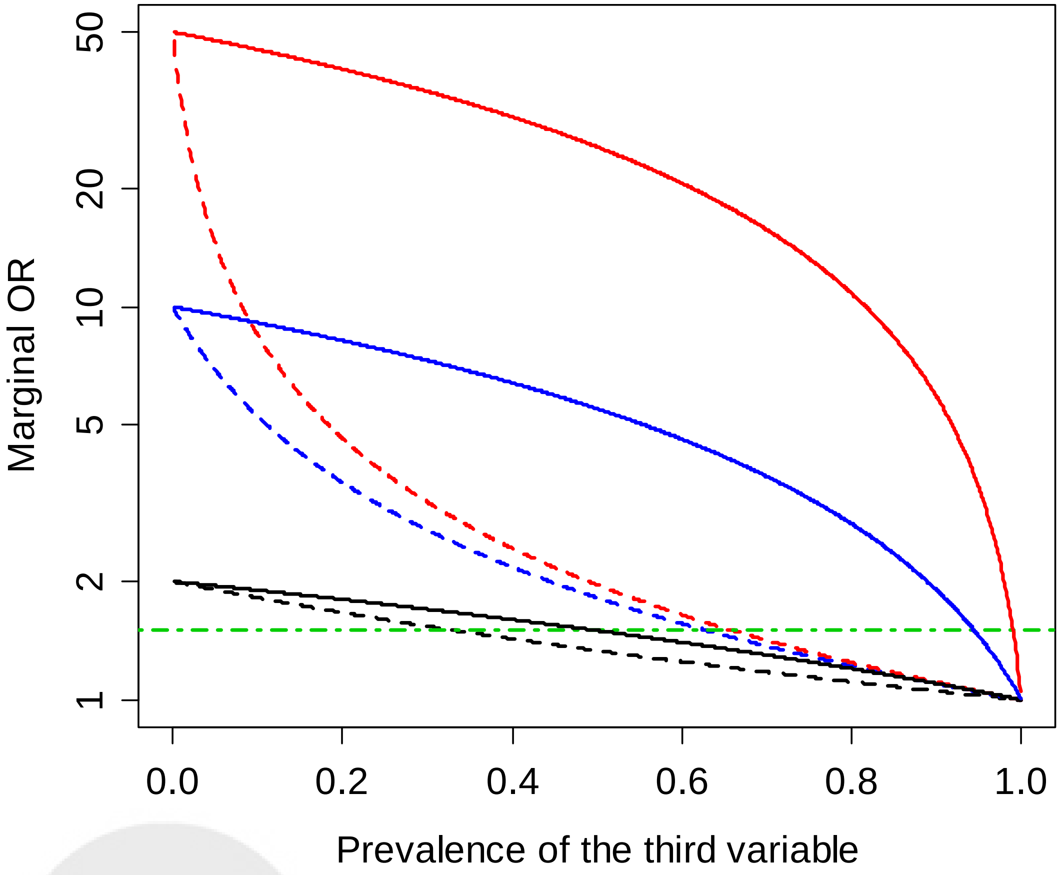

Figure 1. Marginal ORs of SNP-disease associations under the epistasis/interaction of redundancy/masking. The conditional ORs of SNP-disease associations in the absence of "the third variable" are set to be 2.0 (black), 10.0 (blue), and 50.0 (red). The solid lines show the marginal ORs when "the third variable" itself does not modify disease risk, while the broken lines show the marginal ORs when "the third variable" and the SNP modify disease risk by the same multiplicative factor in the absence of the other. These show that the prevalence of "the third variable" has to be very high for the marginal ORs to become below 1.5 (the green broken-dotted line) as seen in GWAS.

Another example may involve a prevalent factor, genetic and/or non-genetic, as “the third variable” which masks SNPs’ associations with disease risk. In this case, the prevalent factor does not have to be associated with the disease risk so long as it is prevalent. These are illustrated by solid lines in Figure 1. As a specific numerical example, $OR_{1|0} = 10.0$, $odds_{01} / odds_{00} = 1.0$, $\delta = -0.9$, and $p_{01} = \gamma \times p_{00}$ would yield the marginal OR = 1.8, 1.4, and 1.2 for γ = 10, 20, and 50, respectively. The OR of 10.0 for the SNP in the absence of the prevalent factor that is not associated with the disease risk, can be attenuated to the marginal OR that is close to 1.0, depending on the prevalence of the prevalent factor. This example may be more plausible than the previous one above because the previous example puts the majority of the population at higher risk. See Table 1 for a real-data example from a GWAS study of Crohn’s Disease [WTCCC, 2007]. Specifically, in the Crohn’s Disease GWAS data of Wellcome Trust Case Control Consortium, the marginal OR estimate of the disease for the CC genotype of rs6518932 is 1.01. The corresponding conditional OR estimate, given “the third variable” (rs5999715’s genotype) being equal to AA/AC, is 61.80. The strong SNP-disease association in the absence of “the third variable”, indicated by the conditional OR estimate of 61.80, is masked by the prevalent “third variable” (99.4% of the control group is rs5999715=CC) into the marginal OR of 1.01.

Table 1. An illustration of “interaction of masking” with real GWAS data

Overall		Disease		Marginal *OR*
		Case	Control	
SNP rs6518932	1 (CC)	1243	2084	1.01
	0 (TT/TC)	502	851	

Third variable (rs5999715) = 0 (AA/AC)		Disease		Conditional $OR_{1|0}$
		Case	Control	
SNP rs6518932	1 (CC)	206	5	61.80
	0 (TT/TC)	8	12	

Note that, while the discussion above, including Figure 1, Table 1, and the specific numerical examples considered risk-elevating SNP-disease associations only, the same attenuation of marginal ORs occurs for risk-reducing SNP-disease associations: this also applies to Figure 2 and its numerical examples below.

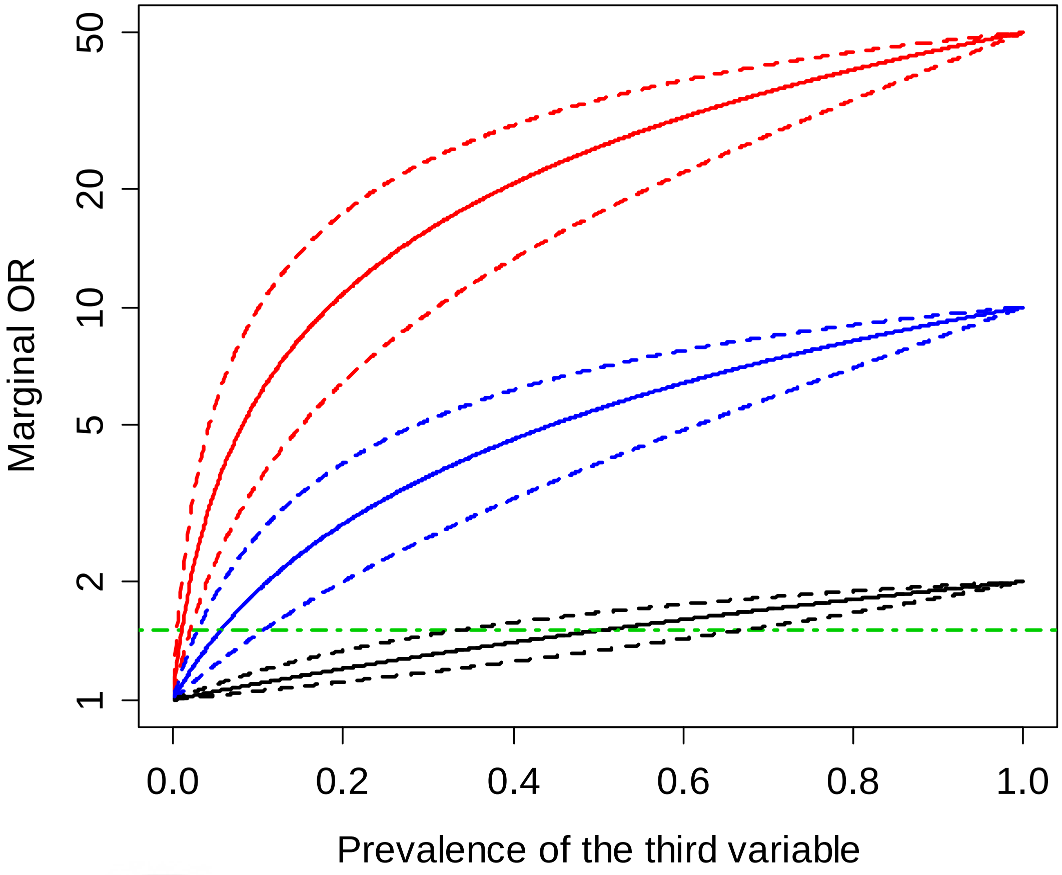

Figure 2. Marginal ORs of SNP-disease associations under the epistasis/interaction of concurrence. The conditional ORs of SNP-disease associations in the presence of "the third variable" are set to be 2.0 (black), 10.0 (blue), and 50.0 (red). The solid lines show the marginal ORs when "the third variable" itself does not modify disease risk in the low-risk genotype group of the SNP: this is the mirror image of the solid lines in Figure 1. The two broken lines above and below each solid line represent the same scenario as the solid line except that "the third variable" increases and decreases, respectively, disease risk by two-fold in the low-risk genotype group of the SNP. The prevalence of "the third variable" has to be less than 5.6% to make the marginal OR less than 1.5,

when the conditional ORs are 10.0 and 1.0 in the presence and absence, respectively, of “the third variable.”

Interaction of Concurrence

This form occurs when the SNP is strongly associated with the disease risk only in the presence of a third factor (i.e., both factors are required). The marginal OR gets close to 1.0 under this scenario if “the third variable” is uncommon (in the SNP=0 group) and/or strongly negatively associated with the disease (in the SNP=0 group). For example, modification of disease risk may require multiple factors, genetic and/or non-genetic, to take place concurrently, similar to the set of multiple mutations required in multistage carcinogenesis for a cancerous cell to develop. These are illustrated in Figure 2. As a specific numerical example, $OR_{1|0} = 1.0 = odds_{01} / odds_{00}$, $OR_{1|1} = 10$ ($\delta = 9$), and $p_{01} = \gamma \times p_{00}$ would lead to the marginal *OR* = 1.8, 1.4, and 1.2 for γ = 1/10, 1/20, and 1/50, respectively, in spite of each factor’s 10-fold modification of disease odds in the presence of the other.

The two forms of interaction are, in fact, equivalent: the same phenomenon can be described in two different ways. For example, masking by a prevalent “third variable” of the SNP-disease association in the absence of “the third variable” can be considered as the concurrence of SNP and the uncommon absence of “the third variable”. Also, one’s risk elevation can be derived from the other’s risk reduction by switching the coding between 0 and 1 for the SNP and that for the third variable. Nevertheless, it is useful to discuss both forms at least initially as the equivalence is not immediately clear.

4. CONCLUSION

In summary, the statistical reasoning above shows that the “missing heritability” problem may be solved if the specific form of interaction, namely, the interaction of concurrence with relatively uncommon genetic and non-genetic factors, is fully explored, assuming that heritability information itself is not missing in GWAS data. The specific form of interaction has been largely unexplored in GWAS where the standard data analysis examines each single SNP one at a time. In particular, if the etiology of the disease is known to have substantial genetic components and it is explained little by GWAS findings, exploration of interaction of concurrence with relatively uncommon genetic and non-genetic factors, may recover the “missing heritability” and identify genetic components with appreciable ORs. While existing tools such as logic regression [Ruczinski *et al.*, 2003] may be useful for such explorations of GWAS data, it would be advantageous if tools are developed specifically targeting the interaction of concurrence with relatively uncommon genetic and non-genetic factors.

5. SUPPLEMENTARY MATRIALS

APPENDIX A: Proof of the theorem for GWAS scenario

Let p_{ij} and $odds_{ij}$ be the proportion among controls and the odds of disease, respectively, with the i^{th} genotype of the SNP of interest (*i* = 0, 1) and the j^{th} category of "the third variable" (*j* = 0, 1, …, *J*). The marginal *OR* of the disease comparing the SNP genotype 1 vs. 0, ignoring the third variable, and the corresponding conditional OR $OR_{1|j}$ in the j^{th} category of "the third variable" are given by:

$$OR = \frac{(\sum_{j=0}^{J} p_{1j} \times odds_{1j}) / \sum_{j=0}^{J} p_{1j}}{(\sum_{j=0}^{J} p_{0j} \times odds_{0j}) / \sum_{j=0}^{J} p_{0j}}, \quad OR_{1|j} = odds_{1j} / odds_{0j}.$$

The condition of the theorem is separated into:

(1) Non trivial case where *the SNP genotype is independent of the third variable in both cases and controls*; and

(2) Trivial case where *the third variable is independent of the case-control status in both SNP-genotype groups.*

Proof of sufficiency

(1) Non trivial case

If the SNP and "the third variable" are independent in controls, then $p_{1j} \times p_{00} = p_{0j} \times p_{10}$ for all *j*.

If the SNP and "the third variable" are independent in cases, then $(p_{1j} \times odds_{1j}) \times (p_{00} \times odds_{00}) = (p_{0j} \times odds_{0j}) \times (p_{10} \times odds_{10})$ for all *j*.

Therefore, $odds_{1j} / odds_{0j} (= OR_{1|j}) = odds_{10} / odds_{00} (= OR_{1|0})$ for all *j*, and also

$$OR = \frac{(\sum_{j=0}^{J} p_{1j} \times odds_{1j}) / \sum_{j=0}^{J} p_{1j}}{(\sum_{j=0}^{J} p_{0j} \times odds_{0j}) / \sum_{j=0}^{J} p_{0j}} = OR_{1|j} \text{ for all } j.$$

(2) Trivial case

If "the third variable" and the disease are independent in both SNP genotype groups, then $odds_{1j} / odds_{10} = odds_{0j} / odds_{00} = 1$ which implies $OR_{1|j} = OR_{1|0} = OR$.

Proof of necessity

If $OR = OR_{1|j} (= odds_{1j} / odds_{0j})$ for all *j*, then

$$OR = \frac{(\sum_{j=0}^{J} p_{1j} \times odds_{1j}) / \sum_{j=0}^{J} p_{1j}}{(\sum_{j=0}^{J} p_{0j} \times odds_{0j}) / \sum_{j=0}^{J} p_{0j}} = OR \times \frac{(\sum_{j=0}^{J} p_{1j} \times odds_{0j}) / \sum_{j=0}^{J} p_{1j}}{(\sum_{j=0}^{J} p_{0j} \times odds_{0j}) / \sum_{j=0}^{J} p_{0j}}$$

and, therefore, $(\sum_{j=0}^{J} p_{1j} \times odds_{0j}) / \sum_{j=0}^{J} p_{1j} = (\sum_{j=0}^{J} p_{0j} \times odds_{0j}) / \sum_{j=0}^{J} p_{0j}$. This is satisfied if and only if:

either $odds_{0j} = odds_{00}$ or $p_{1j} / \sum_{j=0}^{J} p_{1j} = p_{0j} / \sum_{j=0}^{J} p_{0j}$ for all *j*, that is, either

(2) Trivial case holds because $odds_{0j} = odds_{00}$ for all j which implies $odds_{1j} = odds_{10}$ from $OR = OR_{1|j} (= odds_{1j} / odds_{0j})$, and therefore "the third variable" is independent of the case-control status in both SNP-genotype groups; or

(1) Non-trivial case holds because $p_{1j} / \sum_{j=0}^{J} p_{1j} = p_{0j} / \sum_{j=0}^{J} p_{0j}$ for all j, implying the independence of the SNP and "the third variable" in controls ($p_{1j} / p_{0j} = \sum_{j=0}^{J} p_{1j} / \sum_{j=0}^{J} p_{0j}$), and the same applies to cases because of $OR = OR_{1|j} (= odds_{1j} / odds_{0j})$, which implies ($p_{1j} \times odds_{1j} / (p_{0j} \times odds_{0j}) = \sum_{j=0}^{J} (p_{1j} \times odds_{1j}) / \sum_{j=0}^{J} (p_{0j} \times odds_{0j})$).

APPENDIX B: Marginal OR ignoring "the third variable" when the condition of the theorem is not met in cases

For ease of the discussion without loss of generality, we will consider the scenario with a binary third variable (j=0, 1). Based on the Mendelian Randomization, the SNP and "the third variable" are independent in controls, i.e., $p_{11} \times p_{00} = p_{01} \times p_{10}$. Let $OR_{1|0} = odds_{10} / odds_{00}$ denote the conditional OR of the disease comparing the SNP genotype 1 vs. 0, in the category 0 of the third variable. Then, the marginal OR ignoring "the third variable" is given b

$$
\begin{aligned}
OR &= \frac{(p_{11} \times odds_{11} + p_{10} \times odds_{10}) / (p_{11} + p_{10})}{(p_{01} \times odds_{01} + p_{00} \times odds_{00}) / (p_{01} + p_{00})} \\
&= \frac{(p_{11} \times odds_{11} + p_{10} \times odds_{10}) / \dfrac{p_{10}}{p_{00}}}{(p_{01} \times odds_{01} + p_{00} \times odds_{00})} \\
&= \frac{(p_{01} \times odds_{11} + p_{00} \times odds_{10})}{(p_{01} \times odds_{01} + p_{00} \times odds_{00})} \\
&= \frac{(p_{01} \times OR_{1|0} \times odds_{01} + p_{00} \times OR_{1|0} \times odds_{00})}{(p_{01} \times odds_{01} + p_{00} \times odds_{00})} + \frac{\delta(p_{01} \times OR_{1|0} \times odds_{01})}{(p_{01} \times odds_{01} + p_{00} \times odds_{00})} \quad \text{if } OR_{1|1} = OR_{1|0}(1+\delta) \\
&= OR_{1|0} \times \left\{ 1 + \delta \Big/ \left(1 + \frac{p_{00} \times odds_{00}}{p_{01} \times odds_{01}} \right) \right\}.
\end{aligned}
$$

REFERENCES

Clayton, D., and McKeigue, P.M. (2001), “Epidemiological methods for studying genes and environmental factors in complex diseases,” *Lancet*, 358(9290), 1356-1360.

Cornfield, J. (1951), “A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix,” *J Natl Cancer Inst*, 11(6), 1269-1275.

Ducharme, G.R., and LePage, Y. (1986), “Testing collapsibility in contingency tables,” *Journal of the Royal Statistical Society (Series B)*, 48, 197-205.

Goldstein, D.B. (2009), “Common genetic variation and human traits,” *N Engl J Med*, 360(17), 1696-1698.

Gordis, L. (2008), *Epidemiology* (4nd ed.), Philadelphia, PA: W. B. Saunders Company.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F., McCarroll, S.A., and Visscher P.M. (2009), “Finding the missing heritability of complex diseases,” *Nature*, 461(7265), 747-753.

Ruczinski, I., Kooperberg, C., and LeBlanc, M.L. (2003), “Logic Regression,” *Journal of Computational and Graphical Statistics*, 12(3), 475-511.

Simpson, E.H. (1951), “The interpretation of interaction in contingency tables,” *Journal of the Royal Statistical Society (Series B)*, 13, 238–241.

Whittemore, A.S. (1978), “Collapsibility of multidimensional contingency tables,” *Journal of theRoyal Statistical Society (Series B)*, 40, 328–340.

WTCCC (2007). “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, 447, 661-678.