How Can We Explain Very Low Odds Ratios in GWAS? I. Polygenic Models

Susan E. Hodge David A. Greenberg

Battelle Center for Mathematical Medicine, The Research Institute, Nationwide Children's Hospital, and Department of Pediatrics, The Ohio State University, Columbus, OH, USA

Keywords

Case-control association analysis · Complex disorder · Genetic heterogeneity · Genome-wide scan · Linkage disequilibrium · Statistical genetics · Polygenic inheritance · Odds ratios · GWAS

Abstract

Genome-wide association studies of common diseases often identify a number of disease-related SNPs that reach highly significant p values but at the same time show very low disease odds ratios (ORs), most <1.5 and many <1.2 . Despite their statistical significance, associations involving very low ORs explain little about the genetic contribution to the disease and nothing about disease inheritance. A commonly accepted explanation for very low ORs involves a model of polygenic inheritance, i.e., where the disease being studied is caused by a large number of interacting genes, each gene contributing only a small increment to disease risk. Here we demonstrate the perhaps counterintuitive result that, within a reasonable range of disease population prevalences ($\leq 10\%$), a pure polygenic model is incompatible with very low ORs, unless very large numbers (hundreds or even thousands) of polygenic loci are involved.

© 2017 The Author(s)
Published by S. Karger AG, Basel

Introduction

Genome-wide association studies (GWAS) of common diseases often identify a number of disease-related SNPs that reach highly significant p values but at the same time show very low disease odds ratios (ORs), most <1.5 and many <1.2 . For example, 36 of the 39 associations reported in Baxter and Jordan's type 1 diabetes review article [1] have $OR \leq 1.2$ (see Table 2 in Baxter and Jordan's study). A recent GWAS involving thousands of patients and controls [2] identified 128 independent associations. They did not report the actual ORs, but Sekar et al. [3] followed up those markers with the most significant p values and reported relative risks ranging from 1.00 to 1.27 for these loci. In a 2014 GWAS and meta-analysis of breast cancer [4], the highest OR was 1.26. Despite their statistical significance, these associations involving very low ORs explain little about the genetic contribution to the disease and nothing about the mode of inheritance or the possible role of genetic heterogeneity. A commonly accepted explanation [e.g., 2, 5, 6] for very low ORs is that the disease being studied is caused by a large number of interacting genes, each gene contributing only a small, additive increment to disease risk, i.e., polygenic inheritance. Others [e.g., 7] have questioned the automatic as-

sumption that the polygenic model is the most appropriate one.

In this study we have quantified the relationship between population prevalence and OR for pure polygenic models. We determined that within a reasonable range of disease population prevalences ($\leq 10\%$), a pure polygenic model cannot yield very low ORs, unless literally hundreds of high-allele-frequency polygenic genes are involved. We analyze how the interplay of parameters (number of loci, threshold for affectedness, and gene frequencies) explains these apparently counterintuitive results.

We conclude that the low ORs observed from GWAS, even in those genes identified with high statistical significance, cannot be explained simply by assuming a polygenic-type additive interaction, except by invoking very large numbers of loci. One could modify the polygenic model, for example, by allowing unequal allele frequencies (at the disease loci) or dominance. However, such modifications result in higher ORs, since they move the models more toward major gene effect models and away from the polygenic one. We advocate a search for alternative explanations of very low ORs, besides a polygenic model. In particular, and drawing on experience with actual disease, we suggest that researchers pursue the influence of genetic heterogeneity on our understanding of complex diseases.

The Model

Assumptions

We posit a polygenic model as in Falconer [8] and Reich et al. [9]. There are n “disease susceptibility” loci, each with 2 alleles, **A** and **a**. Each **A** contributes 1 “unit” of risk or liability to the individual. For example, if an individual is homozygous **AA** at 1 disease locus and heterozygous **Aa** at 5 additional disease loci, that person’s total liability is 7 units. There exists a threshold k such that individuals with k or more liability units (i.e., individuals with $\geq k$ **A** alleles) are affected, and those with fewer than k are not.

Each **A** allele has the same population frequency p . For convenience, let G_i denote genotype **AA**, **Aa**, or **aa**, for $i = 2, 1, 0$, respectively, and let π_i denote the population frequency of the i -th genotype:

$$\pi_2 = p^2, \pi_1 = 2pq, \pi_0 = q^2, \text{ with } q = 1 - p. \quad (1)$$

From these assumptions and the 3 parameters – n , k , and p – we derive expressions for the population prevalence, the “per-locus penetrance,” and the “proportion with genotype G_i ,” and from them, for the OR itself.

Prevalence

The *population prevalence* represents the probability that an individual has at least k **A** alleles among the disease loci. Since there are n disease loci, there are $2n$ “slots” in which an **A** allele can occur. For convenience, let $m = 2n$. Each locus is assumed independent, and each slot represents an independent Bernoulli trial, with probability p of containing an **A** as opposed to an **a**. Let X represent the number of **A** alleles carried by an individual. Then X has a binomial probability distribution:

$$\begin{aligned} \text{Prev} &= P[\text{individual is affected}] = P[X \geq k] \\ &= \sum_{x=k}^m \binom{m}{x} p^x q^{m-x}, \text{ with } m = 2n. \end{aligned} \quad (2)$$

Per-Locus Penetrance

Our calculations are based on having found susceptibility loci in an association analysis. Let “ L ” represent one of these “loci of interest.” The *per-locus penetrance* f_i answers the question, if an individual is known to have genotype G_i at any given susceptibility locus L , what is the probability that the person is affected? Define $f_i \equiv P[\text{individual is affected} | \text{individual has } G_i \text{ at } L]$, for $i = 0, 1, 2$, or, in a more convenient notation, $f_i = P[\text{affected} | G_i \text{ at } L]$, i.e., $f_i = P[\text{affected} | \text{has } i \text{ **A** alleles at } L]$. This equals the probability that the individual has at least $k - i$ **A** alleles in the remaining $m - 2$ slots (among the n susceptibility loci). Thus,

$$f_i = \sum_{x=k-i}^{m-2} \binom{m-2}{x} p^x q^{m-2-x}, \text{ for } i = 0, 1, 2. \quad (3)$$

Using the π_i from equation 1, it also follows that

$$\sum_{i=0}^2 f_i \pi_i = \text{Prev}. \quad (\text{cf. equation 2})$$

“Proportion with genotype G_i ”

The *proportion with genotype G_i* , which we denote by φ_i , is defined as the probability that an individual has genotype G_i at L , given that the individual is affected: $\varphi_i \equiv P[G_i \text{ at } L | \text{affected}]$. From Bayes’ rule

$$\begin{aligned} \varphi_i &\equiv P[G_i \text{ at } L | \text{affected}] \\ &= \frac{P[\text{affected} | G_i \text{ at } L] P[G_i \text{ at } L]}{P[\text{affected}]} = \frac{f_i \pi_i}{\sum_{j=0}^2 f_j \pi_j} = \frac{f_i \pi_i}{\text{Prev}}. \end{aligned} \quad (4)$$

Odds Ratios

For an association study, one creates 2×2 tables, comparing the proportions that have ≥ 1 **A** allele with the proportion that have no **A** alleles, compared between cases and controls. Among affected individuals (cases), the expected proportion to have at least 1 **A** allele is

$$\varphi_A \equiv \varphi_2 + \varphi_1, \quad (5)$$

from equation 4. For the controls, this expected proportion is simply the population probability of having the **AA** or **Aa** genotype, which equals $1 - q^2$. (This definition assumes that “unscreened”

¹ Each per-locus penetrance f_i in equation 3 can also be construed as the population prevalence of a polygenic disease with the same allele frequency p , but with $n - 1$ loci and threshold $k - i$.

Table 1. Results for a polygenic model with $n = 20$

k	Prev	f_2	f_1	f_0	φ_A	OR	H
<i>Gene frequency $p = 0.10$</i>							
7	0.100	0.323	0.175	0.080	0.349	2.29	1.75
8	0.042	0.175	0.080	0.032	0.385	2.67	2.00
9	0.016	0.080	0.032	0.011	0.421	3.10	2.25
<i>Gene frequency $p = 0.20$</i>							
12	0.088	0.216	0.122	0.062	0.544	2.12	1.50
13	0.043	0.122	0.062	0.029	0.574	2.39	1.62
14	0.019	0.062	0.029	0.012	0.603	2.70	1.75

f_2, f_1 , and f_0 are defined in equation 3, φ_A in equation 5, the OR in equation 6, and $H = k/(mp)$, where $m = 2n$.

Table 2. Results for a polygenic model with $n = 100$

k	Prev	f_2	f_1	f_0	φ_A	OR	H
<i>Gene frequency $p = 0.10$</i>							
26	0.100	0.188	0.134	0.092	0.259	1.49	1.30
28	0.043	0.092	0.061	0.039	0.274	1.61	1.40
30	0.016	0.039	0.024	0.014	0.289	1.73	1.50
<i>Gene frequency $p = 0.20$</i>							
48	0.094	0.148	0.112	0.082	0.441	1.40	1.20
50	0.049	0.082	0.060	0.042	0.454	1.48	1.25
52	0.024	0.042	0.029	0.020	0.467	1.56	1.30

f_2, f_1 , and f_0 are defined in equation 3, φ_A in equation 5, the OR in equation 6, and $H = k/(mp)$, where $m = 2n$.

population controls are used, see Unscreend versus Screend Controls.) Thus, the OR is

$$\text{OR} = \frac{\varphi_A \times q^2}{(1 - \varphi_A) \times (1 - q^2)}. \quad (6)$$

Results

Magnitudes of ORs

Tables 1 and 2 show results for $n = 20$ and 100 loci, respectively. For each n , we consider gene frequencies $p = 0.10$ and 0.20, and for each n - p combination we examine threshold values (k) that lead to prevalences between roughly 1 and 10%. For each model, the tables give the prevalence, the 3 per-locus penetrances (f_i), the proportion of affected individuals with ≥ 1 A allele (φ_A), and the corresponding OR. The rightmost column gives the “threshold ratio” H , see Threshold Ratio H below.

Example

Illustrate with a polygenic disease involving $n = 20$ susceptibility loci ($m = 40$), threshold $k = 8$, and A’s frequency $p = 0.1$ ($q = 0.9$) at each locus. From equation 2, the prevalence of the disease is Prev = 0.042, the per-locus penetrances are $f_2 = 0.175$, $f_1 = 0.080$, and $f_0 = 0.032$, from equation 3. From equation 4, $\varphi_2 = 0.042$, $\varphi_1 = 0.344$, and $\varphi_0 = 0.615$. Thus, the probability that the individual has at least 1 A at locus L is $\varphi_A = 38.6\%$, from equation 5. The OR is

$$\text{OR} = \frac{0.386 \times 0.81}{0.614 \times 0.19} = 2.67. \quad (\text{from equation 6})$$

Tables 1 and 2 show that over the range of parameters considered, it is difficult to achieve ORs below 1.5, and ORs as low as 1.2 are almost impossible. In order to low-

er the OR, the prevalence must be raised to high levels and vice-versa.

Threshold Ratio H

We define a useful quantity, the “threshold ratio” $H = k/(mp)$, where k , the threshold, is the minimum number of A alleles for an individual to be affected, and $mp = E[X]$ is the average number of As that a random individual would carry (recall $m = 2n$). Thus, H is a kind of scaling parameter. Tables 1 and 2 show values of H in the rightmost column. For those combinations of n , p , and k shown in the tables, the threshold ratio falls consistently between about 1.2 and at most 2.25.

Figure 1 illustrates the relationship between the threshold ratio and the corresponding OR for $n = 20, 50, 100$ and $p = 0.1, 0.2$. The figure plots OR versus H for $0.5 \leq H \leq 1.5$. The graphs reveal that ORs below 1.2, the values most frequently found in GWAS analyses, occur only for H values slightly below about 1.0, for $n = 100$, and even lower for smaller n . However, such low values of H correspond to disease prevalences well over 50%, as seen in Figure 2, which plots disease prevalence versus H . For all n - p combinations, $H = 1$ corresponds to $\sim 50\%$ disease prevalence, and $H < 1$ corresponds to even greater prevalences. The 2 figures together give visual expression to the patterns in Tables 1 and 2, showing why it is difficult to achieve simultaneously a very low OR and a reasonable prevalence with the polygenic model.

Extremes

We investigated what would happen if even greater numbers of disease loci, n , were involved in disease expression. With prevalence fixed at approximately 5%, we examined increasing values of n (100, 200, ...) and allele

Fig. 1. Odds ratio (OR) for a polygenic disease plotted vs. threshold ratio, $H = k/(2np)$, for indicated n - p combinations. The dashed line indicates OR = 1.2.

Fig. 2. Prevalence for a polygenic disease plotted vs. threshold ratio, $H = k/(2np)$, for indicated n - p combinations. The dashed line indicates Prevalence = 10%.

Table 3. Minimum possible OR (MPOR) as a function of n for a population prevalence of $\sim 5\%$

a Detailed results for $n = 100$

n	p	k	Prev	OR
100	0.1	28	0.043	1.61
	0.2	50	0.049	1.48
	0.3	71	0.054	1.44
	0.4	92	0.049	1.47
	0.5	112	0.052	1.50

b Summary results for $n = 100, \dots, 500$, and 1,000

n	p	k	Prev	MPOR
100	0.3	71	0.054	1.44
200	0.3	136	0.047	1.31
	0.4	177	0.047	1.31
300	0.3	199	0.051	1.24
	0.4	260	0.052	1.24
400	0.4	343	0.053	1.20
500	0.3	324	0.053	1.18
	0.4	426	0.050	1.18
1,000	0.3	634	0.052	1.12
	0.4	837	0.048	1.13

frequencies $p = 0.10, 0.20, \dots$. For each n - p combination, we set the threshold k at the value corresponding most closely to 5% prevalence. Table 3 shows the results. At first, for given n , the OR decreases with increasing p ; however, at some point (around $p = 0.3$ – 0.4), the OR begins to increase. Thus, for each n we could identify an approximate *minimum possible OR*. Table 3a illustrates these patterns in detail for $n = 100$; Table 3b summarizes results for values of n up to 1,000. In order to come even close to an OR as low as 1.2 and still maintain disease prevalence of $\approx 5\%$, at least 300–400 loci are required, with high gene frequencies p and thresholds k of ≥ 200 . For example, one model that achieves OR = 1.24 involves 300 susceptibility loci, each with a disease allele frequency of 0.3 and a threshold $k = 199$. At least 199 among the 600 “slots” would need to carry an A allele in order for an individual to be affected (Table 3b). To achieve ORs closer to 1.1 would require n in the range of 1,000. These parameter values seem unlikely. Note also that for prevalences $< 5\%$, the results are even more extreme (see Conclusions).

Discussion

The Threshold Ratio

Parameterizing the polygenic models in terms of the threshold ratio $H = k/(mp)$ reveals insights that could otherwise be obscured by the wide variance among the values

Table 4. Standard error (σ) and $1 + 1.28\sigma$ for indicated polygenic models

n, p	$\sigma = \sqrt{\frac{1-p}{mp}}$	$1 + 1.28\sigma$	Value of H at which Prev = 0.1
20, 0.1	0.4743	1.61	off scale
20, 0.2	0.3162	1.41	1.48
50, 0.1	0.3000	1.38	1.45
50, 0.2	0.2000	1.26	1.29
100, 0.1	0.2121	1.27	1.30
100, 0.2	0.1414	1.18	1.20

of n , p , and k . The value of mp represents the average number of **A** alleles in individuals in the general population, so if that number were set as the threshold, about half the population would be affected. This is why $H = 1$ (i.e., $k = mp$) corresponds to prevalence of $\approx 50\%$. Thus, H provides a way to scale the threshold relative to that average number. Figure 2 illustrates this graphically and also shows how the prevalence decreases monotonically with increasing H .

It is also useful to interpret H as a random variable:

$$H = \frac{X}{mp}.$$

Recall that X , the number of **A** alleles that an individual has, is a binomial random variable with mean mp and variance mpq , so the normal approximation for the distribution of H is

$$N\left(1, \frac{q}{mp}\right).$$

The disease prevalence corresponds to 1 minus the cumulative distribution function of H , yielding a prevalence of approximately 50% at $H = 1$ for all n - p combinations. The prevalence range of 10% or less corresponds to $H \geq \mu + 1.28\sigma$ in the normal approximation, with $\mu = 1$ and

$$\sigma = \sqrt{q / (mp)}.$$

Table 4 gives values of σ and $1 + 1.28\sigma$ for the normal approximations corresponding to the polygenic models in Figure 2. These values correspond closely to the values of H at which the curves in Figure 2 cross the Prev = 10% line.

Figure 1 plots the OR versus H , and here the relationship is monotonically *increasing*. In other words, lower prevalences require higher values of H , which in turn correspond to larger ORs.

In summary, in order for the OR to be 1.2 or less, H must be below 1, and generally well below 1, whereas for disease

prevalences to be 10% or lower, H must be well above 1. These contradictory requirements hold over a wide range of parameter values in the polygenic model and show why the polygenic model is not compatible with observed data (except for very high values of n , see Extremes).

Why Are the ORs Not Smaller?

At first it seems intuitively reasonable that a polygenic model, defined in terms of many loci, each with small effect, should yield very low ORs, so our results appear puzzling. The explanation lies in the distinction between “having small effect” and “being one of the loci involved.” It is true that each locus alone has only a small effect on disease risk. But since so many susceptibility alleles are required in order to be affected, the probability of any given locus being involved is relatively high. That probability, ϕ_A , is on the order of k/n , by the following reasoning: If an affected individual had *exactly* k **A** alleles, the probability for any particular slot to contain an **A** would be exactly k/m , and the probability that at least 1 of 2 “slots” in a genotype would carry an **A** would be approximately $2k/m$, which equals k/n . This reasoning does not allow for the possibility of *both* slots in a genotype carrying an **A**, or for the possibility (negligible for small p) that the affected person has $>k$ **A** alleles. However, k/n does provide an approximate magnitude for ϕ_A . For example, with $n = 20$ and $k = 9$, k/n equals 0.45, versus the actual value $\phi_A = 0.421$ for $p = 0.1$ (Table 1). See the Appendix for rigorous derivations.

Unscreened versus Screened Controls

We have defined the OR (equation 6) using controls that are selected at random from the general population. Alternatively, one could screen the population sample and use as controls only those individuals who are not affected with the disease being studied [10]. However, we did not investigate ORs using screened controls, because those ORs are always *greater* in magnitude than formula 6.² For example, if controls were screened in the illustra-

² The proportion of unaffected individuals in the population who carry the **A** allele at a given locus is

$$\begin{aligned}\Psi_A &\equiv P[L \text{ has } \mathbf{AA} \text{ or } \mathbf{Aa} | \text{unaffected}] \\ &= 1 - \frac{(1-f_0)\pi_0}{1-\text{Prev}} = 1 - \left(\frac{1-f_0}{1-\text{Prev}}\right)q^2,\end{aligned}$$

and the corresponding OR is

$$\text{OR}_{\text{screened}} = \frac{\phi_A \times (1 - \Psi_A)}{(1 - \phi_A) \times \Psi_A}.$$

The OR in equation 6 is lower in magnitude than $\text{OR}_{\text{screened}}$, since the former includes affected individuals in the controls, thus diluting the apparent difference between the cases and controls.

Table 5. Effect of unequal allele frequencies on ORs: some examples with $n = 20$ and $k = 8$

Allele frequency		OR	
p	r	Locus 1	Loci 2–20
0.10	0.10	2.67	2.67
0.10	0.09	3.00	2.96
0.10	0.05	5.95	5.23
0.90	0.05	18.89	4.14

p is the frequency of the **A** allele at locus 1; r is **A**'s frequency at each of the remaining 19 loci. Row 1, with $p = r = 0.1$, is identical to the example with $p = 0.1$ and $k = 8$ in Table 1. In rows 2 and 3, p remains constant, while r is decreased. Row 4 shows a more extreme example in which locus 1 has become essentially a “major” locus. For rows 2–4, not only does the OR for locus 1 increase, but so does the OR for each of the remaining loci, compared to when p and r are equal.

tion in the section Magnitudes of ORs, the OR would be 2.83, as compared to OR = 2.67 seen there. Since we are trying to determine how to achieve *low* ORs in this study, we did not investigate ORs based on screened controls.

Possible Variations on the Model

1. In the Assumptions section, we stipulated that each **A** allele contributes 1 unit of risk. Alternative models could, e.g., stipulate that each **A** contributes r units of risk ($r < 1$) and/or that **a** alleles have protective effects *against* disease. These kinds of variants would represent reparameterizations of the model we explore here, so the results would be similar, but expressed in different parameters.

2. By setting the disease allele frequencies to be the same at all n susceptibility loci, we have created a “purely” polygenic model. If the **A** allele were more frequent at some loci than at others (i.e., p not equal for all n disease loci), the model would move more in the direction of major locus models and ORs would *increase*. For example, consider a model with $n = 20$ susceptibility loci and threshold $k = 8$, where locus 1 has allele frequency p , whereas loci 2 through 20 each have the same allele frequency r , with $r < p$. Table 5 shows a few selected examples, indicating that as r diverges farther from p , the ORs increase. This applies both to the OR calculated at locus 1 *and* to an OR calculated at any one of the other 19 loci. Thus, by setting all allele frequencies equal we have created the *best* case for achieving very low ORs, within the polygenic framework.

3. Our model formulation assumes that L , the locus being studied, is not merely a marker, but is itself one of the disease susceptibility loci, or else is in very tight linkage disequilibrium (LD) with one of the susceptibility loci. If that is not the case, i.e., if L is in only loose LD with the disease locus, this will attenuate the OR. However, this circumstance would severely affect the utility and meaning of the association itself.

Conclusions

We have investigated the “pure” polygenic model, defined as one with many disease susceptibility loci, each contributing a small, equal, additive amount to disease susceptibility. We have shown that for population prevalences under 10%, this model is not compatible with ORs as low as those commonly observed in GWAS (unless hundreds of susceptibility loci, each with high gene frequency, are involved; see Extremes section). One can also see from Figures 1 and 2 that the situation is not much better for, e.g., prevalence = 20%. Model modifications such as allowing unequal allele frequencies at the disease loci move more toward major gene effect models and away from the polygenic one. In such cases, the major gene(s) will predominate, and the gene(s) in question should be easily identifiable, with high ORs as well as significant p values. However, that is not what has been reported in the majority of GWAS studies of common disease.

Some researchers have argued that very large numbers ($n =$ several hundred or a thousand) of polygenic loci, such as those in Table 3, are realistic. For example, Ott [11] presents a potential polygenic model for schizophrenia. He shows that for $n = 100$ loci (our $n =$ his m), allele frequency $p = 0.0242$ (our $p =$ his q), and threshold $k = 11$, the model matches the population prevalence of ~ 0.01 for schizophrenia. Using those parameter values, he derives segregation ratios for **AA**, **AU**, and **UU** mating types (**A** = affected, **U** = unaffected) and compares them favorably with observed segregation ratios for schizophrenia. However, his study focuses on risk prediction, not ORs, as we do in this paper. Consequently, we calculated the OR for Ott's example. This yielded OR = 2.52, which is much higher than reported ORs for schizophrenia [e.g., 3]. We also note that *any* specified prevalence can be made compatible with numerous combinations of n , p , and k . To illustrate, we examined a number of such parameter combinations for the prevalence of 1% from Ott's example, with thresholds (k) ranging from 8 to 20 for $n =$

50, from 10 to 20 for $n = 100$, and from 10 to 100 for $n = 200$, etc. The minimum possible OR from these models was 1.43 (which occurred for $n = 200$, $k = 100$, $p = 0.201$). This is not surprising: in Table 3 we see that very large values of n are required to achieve very low ORs with prevalence = 5%; this is much more so for a prevalence of 1%.

As suggested in previous work [12] and generalized here, gene-gene interaction models are not proving helpful for explaining very low ORs. Loose LD between the locus being studied and the closest disease susceptibility locus could lead to very low ORs, but as mentioned above, this explanation would render the associations almost meaningless. Another way to account for very low ORs would be to specify reduced “penetrance” within the polygenic model itself, e.g., to let $P[\text{affected} | X \geq k]$ be less than 1, instead of equaling 1, as in this study. However, this would mean that there are still other factors affecting disease pathogenesis, and the polygenic model is not really explaining the very low ORs. A third possibility could be genetic heterogeneity. That is, these diseases may indeed be caused by a large number of genes, but, as opposed to each locus contributing a small amount to one disease, these loci would be contributing *major* risks to a number of genetically *distinct* diseases. If this is correct, it will be critical to carefully phenotype affected individuals, so as to tease out distinct forms and analyze them separately – a design to which enormous GWAS datasets do not lend themselves. In future work, we will explore these alternatives further.

Finally, our sense is that polygenic models involving several hundred (or even a thousand) loci, with high thresholds and high gene frequencies, while mathematically possible, do not seem biologically plausible. Others may disagree. Papers such as this one, as well as Ott’s work, highlight the complexity of the question. In any case, Ioannidis’ suggestion that there is enough support for the polygenic model to justify using it for making accurate clinical predictions [13, third paragraph] seems highly premature, and even dangerous, to us. We trust that time will reveal the truth – and that the truth will no doubt differ for different diseases.

Acknowledgments

This project was supported by The Research Institute, Nationwide Children’s Hospital, Columbus, OH, USA.

Appendix

Alternative Derivation of and Approximations for φ_A

First consider an affected individual with *exactly* k **A** alleles among the n susceptibility loci. The probability that this person has ≥ 1 **A** alleles at locus L is

$$\begin{aligned} P[L \text{ has } \geq 1 \text{ A allele} | X = k] \\ = P[\text{maternal locus has an A}] + P[\text{paternal locus has an A}] \\ - P[\text{both loci have A}], \end{aligned} \quad (\text{A1})$$

all conditioned on $X = k$. The first 2 probabilities each equal

$$\frac{k}{m}.$$

The third probability equals $P[\text{maternal locus has A}] \times P[\text{paternal locus has A} | \text{maternal locus has A}]$

$$= \frac{k}{m} \times \frac{k-1}{m-1}.$$

Substituting these probabilities into equation A1 yields

$$\begin{aligned} P[L \text{ has } \geq 1 \text{ A allele} | X = k] \\ = 2\left(\frac{k}{m}\right) - \frac{k}{m} \times \frac{k-1}{m-1} = \frac{k}{n} - \frac{k(k-1)}{m(m-1)}. \end{aligned} \quad (\text{A2})$$

To derive φ_A , condition on the affected individual having *at least* k **A** alleles:

$$\begin{aligned} \varphi_A = P[L \text{ has } \geq 1 \text{ A allele} | X \geq k] \\ = \frac{P[L \text{ has } \geq 1 \text{ A allele} \& X \geq k]}{P[X \geq k]} = \frac{\text{Num}}{\text{Den}}. \end{aligned} \quad (\text{A3})$$

The numerator of equation A3 is found by summing equation A2 over x from k to m , each term weighted by $P[X = x]$:

$$\begin{aligned} \text{Num} &= \sum_{x=k}^m P[L \text{ has } \geq 1 \text{ A allele} | X = x] P[X = x] \\ &= \sum_{x=k}^m \left[\frac{x}{n} - \frac{x(x-1)}{m(m-1)} \right] \binom{m}{x} p^x q^{m-x}. \end{aligned}$$

The denominator of equation A2 is the prevalence, as in equation 2 in the text, so equation A3 becomes

$$\varphi_A = \frac{\sum_{x=k}^m \binom{m}{x} p^x q^{m-x} \left[\frac{x}{n} - \frac{x(x-1)}{m(m-1)} \right]}{\sum_{x=k}^m \binom{m}{x} p^x q^{m-x}}. \quad (\text{A4})$$

The result in equation A4 is exact and yields values identical to those from equations 4 and 5 in the text.

For small values of p , the probability of an affected individual having more than k **A** alleles becomes negligible, and

$$\varphi_A \approx \frac{k}{n} - \frac{k(k-1)}{m(m-1)}, \text{ as in equation A2.}$$

Table A1. Two approximations for ϕ_A , compared to exact values, for selected combinations of k , n , and p

n	k	$\frac{k}{n}$	$\frac{k}{n} - \frac{k(k-1)}{m(m-1)}$	ϕ_A (exact) for indicated p	
20	8	0.400	0.364	0.1: 0.385	0.01: 0.366
50	18	0.360	0.329	0.1: 0.342	0.05: 0.334
50	25	0.500	0.439	0.1: 0.446	0.05: 0.442
100	40	0.400	0.361	0.1: 0.366	0.05: 0.363

Table A1 shows both

$$\frac{k}{n} \text{ and } \frac{k}{n} - \frac{k(k-1)}{m(m-1)}$$

for selected n - k combinations and compares these approximations with some exact values of ϕ_A , as a function of p . This confirms that

$$\frac{k}{n}$$

gives a rough idea of the magnitude of ϕ_A , and

$$\frac{k}{n} - \frac{k(k-1)}{m(m-1)}$$

gives an even better approximation, especially for small values of p .

References

1 Baxter AG, Jordan MA: From markers to molecular mechanisms: type 1 diabetes in the post-GWAS era. *Rev Diabet Stud* 2012;9:201–223.
2 Schizophrenia Working Group of the Psychiatric Genomics Consortium: Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014;24:421–427.
3 Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, Tooley K, Presumey J, Baum M, Van Doren V, Genovese G, Rose SA, Handsaker RE; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Daly MJ, Carroll MC, Stevens B, McCarroll SA: Schizophrenia risk from complex variation of complement component 4. *Nature* 2016;530:177–183.

4 Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, Schmidt MK, Chang-Claude J, Bojesen SE, Bolla MK, et al: Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* 2013;45:353–361.
5 Potenciano V, Abad-Grau MM, Alcina A, Mateszna F: A comparison of genomic profiles of complex diseases under different models. *BMC Med Genomics* 2016;9:3.
6 Gibson G: Rare and common variants: twenty arguments. *Nat Rev Genet* 2015;12:135–145.
7 Génin E, Clerget-Darpoux F: The missing heritability paradigm: a dramatic resurgence of the GIGO syndrome in genetics. *Hum Hered* 2015;79:1–4.
8 Falconer DS: *Quantitative Genetics*. London, Longman Group, 8th printing, chapt 6, 1976.

9 Reich T, Rice J, Cloninger CR, Wette R, James J: The use of multiple thresholds and segregation analysis in analyzing the phenotypic heterogeneity of multifactorial traits. *Ann Hum Genet* 1979;42:371–390.
10 Hodge SE, Subaran RL, Weissman MM, Fyer AJ: Designing case-control studies: decisions about the controls. *Am J Psychiatry* 2012;169:785–789.
11 Ott J: Polygenic models for risk prediction in human genetics. *Hum Hered* 2015;80:162–164.
12 Hodge SE, Hager VR, Greenberg DA: Using linkage analysis to detect gene-gene interactions. 2. Improved reliability and extension to more-complex models. *PLoS One* 2016;11:e0146240.
13 Ioannidis JPA: Making optimal use of and extending beyond polygenic additive liability models. *Hum Hered* 2015;80:158–161.