Journal of Theoretical Biology 403 (2016) 68–74

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Genetic risks and genetic model specification

Gang Zheng [a,1], Wei Zhang [b,1], Jinfeng Xu [c], Ao Yuan [d,*], Qizhai Li [b], Joseph L. Gastwirth [e]

[a] National Heart, Lung and Blood Institute, Bethesda, MD, USA
[b] Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
[c] The University of Hong Kong, Hong Kong
[d] Georgetown University, Washington DC, USA
[e] George Washington University, Washington DC, USA

HIGHLIGHTS

- The relationships between GR and OR at the disease locus are obtained.
- The relationships between GR and OR at the marker locus are obtained.
- The procedures for choosing the genetic model is proposed.

ARTICLE INFO

Article history:
Received 1 December 2015
Received in revised form
12 April 2016
Accepted 6 May 2016
Available online 13 May 2016

Keywords:
Case-control design
Genetic model
Genetic risk
Hardy–Weinberg equilibrium

ABSTRACT

Genetic risks and genetic models are often used in design and analysis of genetic epidemiology studies. A genetic model is defined in terms of two genetic risk measures: genotype relative risk and odds ratio. The impacts of choosing a risk measure on the resulting genetic models are studied in the power to detect association and deviation from Hardy–Weinberg equilibrium in cases using genetic relative risk. Extensive simulations demonstrate that the power of a study to detect associations using odds ratio is lower than that using relative risk with the same value when other parameters are fixed. When the Hardy–Weinberg equilibrium holds in the general population, the genetic model can be inferred by the deviation from Hardy–Weinberg equilibrium in only cases. Furthermore, it is more efficient than that based on the deviation from Hardy-Weinberg equilibrium in all cases and controls.

1. Introduction

In the last several decades, genome-wide association studies (GWAS) have been considered as a big success in searching for the deleterious genetic susceptibilities and hundreds of complex traits have been reported to be associated with the genetic variants, such as obesity (Locke et al., 2015), type 1 diabetes (Todd et al., 2007), type 2 diabetes (Altshuler et al., 2000), carcinoid heart disease (CHD) (Korse et al., 2009). Association studies are a major tool for identifying genes conferring susceptibility to complex traits. These traits are termed complex because they are influenced by both genetic and environmental factors. In the design and analysis of genetic association studies, genetic models are often used. A genetic model refers to Mendel's mode of inheritance (Vogel and Motulsky, 1986). Mendel observed 50% heterozygote *AB* and 25% homozygote of each parental type *AA* or *BB* from the crossing of two heterozygotes $AB \times AB$. He found that the phenotype is not always determined by a genotype, and called the allele that determines the phenotype of *AB* dominant, other allele recessive.

A genetic model is a functional relationship of a risk measure given genotypes. For a binary trait, e.g. case-control data, the risk measure is genotype relative risk (GRR) or odds ratio (OR). It should be noted that for a case-control study, the relative risk can be only approximately estimated for a rare disease under the assumption of Hardy-Weinberg Equilibrium. For a quantitative trait, the risk measure is based on the means of the trait given genotypes. In testing a genetic association, specifying a genetic model is equivalent to specifying an alternative hypothesis. Design and analysis with a correctly specified alternative hypothesis are generally more powerful than those with a more broad alternative hypothesis. On the other hand, if the model is incorrectly specified, the outcomes are not satisfactory. Thus, how to specify a correct genetic model and avoid specifying a wrong genetic model are

* Corresponding author.
E-mail address: yuanao@hotmail.com (A. Yuan).
[1] Co-first authors.

http://dx.doi.org/10.1016/j.jtbi.2016.05.016
0022-5193/

important in both design and analysis of genetic association studies. In genetics literatures, the genetic models are generally classified in term of two measures: genotype relative risk and odds ratio. However, no existing literatures have compared the influence of the two risk measures in the genetic association studies.

The Hardy–Weinberg equilibrium is a very important law in the analysis of genetic data. Assume for a biallelic locus with the allele A and B. Denote the genotypes by $(G_0, G_1, G_2) = (AA, AB, BB)$. The HWE states that under random mating the genotype frequencies in a population satisfy the relations: $\Pr(G_0) = q^2$, $\Pr(G_1) = 2q(1-q)$, $\Pr(G_2) = (1-q)^2$, where q is the allele frequency of A. If there is no disturbing factors, such as population stratification, genotyping error, the HWE law usually hold in the source population.

In this paper, we examine the roles of using risk measures in specifying genetic models and how they influence the design and analysis of genetic association studies. We focus on case-control studies but the quantitative trait is also briefly discussed. Then we study how to specify a genetic model using the information of derivation from Hardy–Weinberg equilibrium in cases.

2. Risk measures and genetic models

2.1. Notation

Consider a disease locus with alleles A and B. Denote the genotypes by $(G_0, G_1, G_2) = (AA, AB, BB)$, the disease prevalence by κ, and the penetrance by $f_i = \Pr(D|G_i)$ $(i = 0, 1, 2)$, where D stands for the disease. In a case-control association study, the counts for G_i in r cases and s controls are denoted by r_i and s_i, respectively $(i = 0, 1, 2)$. Denote $n_i = r_i + s_i$, $i = 0, 1, 2$ and $n = n_0 + n_1 + n_2$.

For a quantitative trait, let the random trait be $T = \mu + g + \epsilon$, where μ is the mean in the absence of the genetic effect, g is the genetic effect with $g = -a, d$ and a (a and d are two constants, $a > 0$ and $-a \le d \le a$) for genotypes G_0, G_1 and G_2, and ϵ is a random error with 0 mean. Denote $\mu_i = \mathrm{E}(T|G_i)$, the conditional mean given the genotype $(i = 0, 1, 2)$.

2.2. Binary traits

The GRRs and ORs are denoted by $\mathrm{GRR}_i = f_i/f_0$ and $\mathrm{OR}_i = \{f_i(1-f_0)\}/\{f_0(1-f_i)\}$ $(i=1,2)$. The null hypothesis is H_0: $\mathrm{GRR}_1 = \mathrm{GRR}_2 = 1$ or H_0: $\mathrm{OR}_1 = \mathrm{OR}_2 = 1$. Genetic models are only relevant under the alternative hypothesis H_1. Without loss of generality, assume B is the risk allele. Using the GRRs, the genetic model is recessive (REC) if $\mathrm{GRR}_1 = 1$, additive (ADD) if $\mathrm{GRR}_1 = (1 + \mathrm{GRR}_2)/2$, and dominant (DOM) if $\mathrm{GRR}_1 = \mathrm{GRR}_2$. The definitions of genetic models using ORs are similar.

The alternative hypothesis H_1 can be specified in terms of a genetic model, including the three common genetic models, using GRRs with x or ORs with y as

$$H_1(x) = \{(\mathrm{GRR}_1, \mathrm{GRR}_2) \neq (1, 1): \mathrm{GRR}_1 = 1 - x + x\mathrm{GRR}_2,\ x \in [0, 1]\}; \quad (1)$$

$$H_1(y) = \{(\mathrm{OR}_1, \mathrm{OR}_2) \neq (1, 1): \mathrm{OR}_1 = 1 - y + y\mathrm{OR}_2,\ y \in [0, 1]\}, \quad (2)$$

The REC, ADD and DOM models correspond to $x=0$, $x = 1/2$ and $x=1$ in $H_1(x)$ and $y=0$, $y = 1/2$ and $y=1$ in $H_1(y)$, respectively. Specifying $x = x_0 \in [0, 1]$ or $y = y_0 \in [0, 1]$ is equivalent to specify an alternative hypothesis $H_1(x_0)$ and $H_1(y_0)$. Then sample size/power and analysis can be done for the specified $H_1(x_0)$ or $H_1(y_0)$ rather than for $H_1(x)$ or $H_1(y)$ with all $x \in [0, 1]$ or $y \in [0, 1]$.

The first question to address is that whether or not the genetic models defined using GRRs and ORs are equivalent. The following properties of GRRs and ORs can be readily verified: (i) $\mathrm{GRR}_2 \ge \mathrm{GRR}_1$ if and only if $\mathrm{OR}_2 \ge \mathrm{OR}_1$ and $\mathrm{GRR}_1 \ge 1$ if and only if $\mathrm{OR}_1 \ge 1$, where all the equalities hold simultaneously; (ii) $\mathrm{OR}_2 > \mathrm{GRR}_2$ and $\mathrm{OR}_1 \ge \mathrm{GRR}_1$, where the equality holds under the REC model; (iii) if $\mathrm{GRR}_1 = (1 + \mathrm{GRR}_2)/2$, then $\mathrm{OR}_1 < (1 + \mathrm{OR}_2)/2$; and (iv) let x and y be given in Eqs. (1) and (2), then $y \le x$, where the equality holds under either REC or DOM models.

Note that (i) implies that the REC or DOM models can be defined in terms of either GRRs or ORs. However, from (iii), the ADD model defined using GRRs differs from that defined using ORs. Further, from (ii), the ORs are always larger than the GRRs given the same genetic model. (iv) further implies, except for the REC and DOM models, any other model defined using the GRRs with x corresponds to a *different* model defined using the ORs with y and that $y < x$. In other words, *only* the REC and DOM models do not depend on the risk measures. Other models between the REC and DOM models are *only* well defined given the risk measure.

Table 1 shows the values of y given $x = 0, 1/2, 1$, minor allele frequencies (MAFs) and κ. We assume that the Hardy–Weinberg Equilibrium (HWE) proportion holds in the source population. It indicates that MAF and κ slightly affect the ORs given the GRRs, but a larger κ would increase the ORs for fixed GRRs. The results in Table 1 have an implication in the design of association studies. For example, under the REC model with MAF=0.3, $\kappa = 0.10$, the power to detect GRR=1.5 is the same as that to detect OR=1.584 with the same sample size. In other words, given the same sample size, the power of a study to detect $\mathrm{OR}_2 = 1.5$ would be lower than that with $\mathrm{GRR}_2 = 1.5$. For illustration, we assume $\kappa = 0.05$, $r = s = 1000$, and a two-sided trend test. The powers to detect $\mathrm{OR}_2 = 1.5$ under the ADD model are 34.9%, 68.9%, 87.2%, and 89.4% for MAF=0.05, 0.15, 0.30, and 0.45, respectively, while the powers to detect $\mathrm{GRR}_2 = 1.5$ under the same genetic model are 41.9%, 75.9%, 90.1%, and 92.5%, respectively. The power differences using GRRs or ORs could be quite substantial.

Both GRRs and ORs are commonly used in the literature for sample size and power calculations (Slager and Schaid, 2001; Freidlin et al., 2002; Jackson et al., 2002; Pfeiffer and Gail, 2003) and for deriving test statistics (Sasieni, 1997; Chen and Chaterjee, 2007). Our results indicate that in design of case-control genetic association studies, one should pay attention to the choice of a risk measure in calculating power and sample size, especially the same risk measure should be used when comparing different study designs and analysis. The GRRs and ORs have a one-to-one relationship given f_0. Given GRR_i, $\mathrm{OR}_i = \mathrm{GRR}_i(1-f_0)/(1-\mathrm{GRR}_i f_0)$ $(i = 0, 1, 2)$. Given OR_i, $\mathrm{GRR}_i = \mathrm{OR}_i/(\mathrm{OR}_i f_0 + 1 - f_0)$ $(i = 0, 1, 2)$. With these formulas, one can convert from GRRs to ORs and vice versa.

We have discussed genetic models and risk measures at a disease locus. In practice, only a marker locus is observed, which is assumed to be in linkage disequilibrium (LD) with the disease

Table 1
ORs for a given genetic model defined by GRRs with $\mathrm{GRR}_2 = 1.5$. The disease prevalence $\kappa = 0.05, 0.10, 0.20$. x and y are two indexes for the GRR and OR, respectively, which are the same as those in Eqs. (1) and (2).

κ	MAF	REC ($x=0$)			ADD ($x = 1/2$)			DOM ($x=1$)		
		y	OR1	OR2	y	OR1	OR2	y	OR1	OR2
0.05	0.05	0	1	1.540	0.494	1.266	1.539	1	1.539	1.539
	0.15	0	1	1.540	0.494	1.265	1.538	1	1.535	1.535
	0.30	0	1	1.539	0.494	1.264	1.535	1	1.532	1.532
	0.45	0	1	1.537	0.495	1.263	1.533	1	1.529	1.529
0.10	0.05	0	1	1.540	0.486	1.285	1.586	1	1.583	1.583
	0.15	0	1	1.587	0.487	1.283	1.581	1	1.576	1.576
	0.30	0	1	1.584	0.488	1.280	1.575	1	1.568	1.568
	0.45	0	1	1.579	0.489	1.278	1.570	1	1.563	1.563