Risk Ratios

In human genetics and genetic epidemiology, risk ratios can assume a number of different forms, including risk ratios for relatives, for candidate genes, and for genetic markers. The goal of many genetic studies is to quantify the risk of disease occurrence associated with particular genetic factors. The strength of this association can depend on interactions between environmental and genetic factors, gene–gene interactions, and the distance along a causal pathway from a genetic variant to a disease outcome. Models that can incorporate these complexities have an important role. Here, however, we focus on risk ratios corresponding to associations between a single disease and a single genetic factor.

Let $D$ denote the disease under study, and let $G$ or $\overline{G}$ denote the presence or absence of a particular genetic characteristic in an individual. In these risk ratios, exposure is defined in terms of an individual’s genetic information $G$. Further assume that every individual is correctly classified as having $(D)$ or not having $(\overline{D})$ the disease. Then, a general genetic relative risk (RR) is defined as

$$RR = \frac{Pr(D | G)}{Pr(D | \overline{G})} \quad (1)$$

This measure of disease-by-genetic-factor association is just the ratio of the conditional probabilities of having the disease given the presence or absence of the genetic characteristic. These conditional probabilities are usually referred to as penetrances, and require cross-sectional or cohort designs for their direct estimation. Individuals or families are often sampled according to a disease-related phenotype, so the odds ratio is also useful because of its invariance to the direction of sampling. The genetic odds ratio (OR), that approximates the genetic RR for a rare disease, is

$$OR = \frac{Pr(D | G)}{Pr(D | \overline{G})} \quad (2)$$

which can be written equivalently as

$$OR = \frac{Pr(G | D)}{Pr(G | \overline{D})} \quad (3)$$

Risk Ratios for Relatives

When a genetic factor cannot be measured directly, but information on disease status is available on family members of an affected individual, genetic risk ratios can be derived indirectly.

**Familial Aggregation**

Evidence for familial aggregation, which is the tendency of disease to cluster in families, provides a rationale for subsequent genetic studies intended to assess particular genetic factors or to search for disease susceptibility genes. A measure of familial aggregation that uses information on family history (FH) of disease, say, in first-degree relatives, is

$$OR = \frac{Pr(D | FH)}{Pr(\overline{D} | FH)} \quad (4)$$

In a case–control design, $D$ and $\overline{D}$ correspond to affected cases and unaffected controls, while FH is a surrogate for genetic loading in the family, and is thus subject to misclassification error [13]. Even in the absence of any genetic etiology, however, the probability of a positive FH increases with the number of relatives considered. Alternatively, FH scores can be defined to take into account the number of affected relatives and the family structure.

In a family-based case–control design, familial aggregation can also be assessed by comparing the risk of disease among relatives of cases with that among relatives of controls,

$$Pr(D \text{ in relative type } R \text{ of case} | \text{affected case}) \quad Pr(D \text{ in relative type } R \text{ of control} | \text{unaffected control}) \quad (5)$$

When this ratio is greater than 1, family aggregation can be present, but, without additional environmental exposure information, aggregation due to shared genes can be indistinguishable from that due to shared environment [5].

**Recurrence Risk**

A related measure is known as the recurrence risk for a type $R$ relative of an affected individual. It occurs in the numerator of the ratio, (5), for familial aggregation in family case–control designs. For controls that are representative of the general population, the denominator of this ratio approximates the population risk.
Risk Ratios for Candidate Genes

Investigations of candidate genes are usually based on a priori biologic hypotheses about a particular candidate gene. If genetic information is available, say in the form of a measured candidate gene, then the association of particular genetic variants with a disease can be evaluated using several versions of the genetic risk ratio. These include allelic, genotype, and haplotype RRs, although the latter is usually considered in the context of a genetic marker (see further below).

A common study design involves sampling individuals by disease status, assembling an appropriate control group, genotyping cases and controls at a candidate gene locus, and then comparing the distribution of the candidate gene between the case and control groups. Population (case–population control) and family-based (case–parental control) designs are two approaches used to assess risk ratios for the association of a candidate gene with a disease phenotype.

In case–control studies of candidate gene loci, a fundamental issue is the choice of a reference or control group. Controls can be randomly selected from the population of unaffected individuals or can be matched to the affected cases on relevant characteristics. Because genetic factors vary greatly by ethnic background and population history and geography, there is a serious potential for confounding by population stratification when unrelated individuals are used as controls or when matching on available measures of ethnicity is inadequate. To avoid this problem, the control group can be drawn from members (or potential members) of the family of a case instead of unrelated individuals. However, population controls may be less costly to recruit, and, in the absence of population stratification, may be more efficient statistically (see Family-Based Case–Control Studies).

Population Risk Ratios

At a candidate gene locus, each individual inherits one genetic variant, known as an allele, from each parent and the two alleles together constitute a genotype. When only two variants occur in a population of individuals, there are three possible genotypes, while for a locus with multiple variants (n alleles), there are n(n + 1)/2 possible genotypes. For a single multiallelic locus, the genetic factor G can be expressed as

\[ G = \begin{cases} a_i & i = 1, \ldots, n, \\ a_ia_j & i, j = 1, \ldots, n, \end{cases} \]

For a candidate gene with two alleles, a1 and a2, we can define genotype relative risks in which G depends on the presence or absence of allele a2, i.e.

\[ RR_1 = \frac{Pr(D|a_1a_2)}{Pr(D|a_1a_1)}, \quad RR_2 = \frac{Pr(D|a_2a_2)}{Pr(D|a_1a_1)} \]

and the corresponding genotype odds ratios as

\[ OR_1 = \frac{Pr(D|a_1a_2)/Pr(D|a_1a_1)}{Pr(D|a_1a_1)/Pr(D|a_1a_1)}, \quad OR_2 = \frac{Pr(D|a_2a_2)/Pr(D|a_1a_1)}{Pr(D|a_1a_1)/Pr(D|a_1a_1)}. \]
Family-Based Risk Ratios with Parental Controls

One particular family-based design involves the ascertainment of an affected individual followed by genotyping of this case and their parents. Falk & Rubinstein [2] proposed that the maternal and paternal alleles transmitted to the affected child form a case genotype (D) while the nontransmitted alleles form a control genotype (D).

For a two-allele candidate gene locus with alleles $a_1$ and $a_2$, a simple genotype relative risk is

$$RR = \frac{Pr(D|\text{presence of allele } a_2)}{Pr(D|\text{absence of allele } a_2)}, \quad (9)$$

which does not distinguish between genotypes with one or two copies of allele $a_2$. When the measured candidate gene is the disease gene, this RR is insensitive to population stratification, and can be estimated without bias by the OR [4, 6]:

$$OR = \frac{Pr(a_2 \text{ is present}|D)/Pr(a_2 \text{ is absent}|D)}{Pr(a_2 \text{ is present}|\overline{D})/Pr(a_2 \text{ is absent}|\overline{D}).} \quad (10)$$

This OR is sometimes referred to as a haplotype relative risk, particularly when a genetic marker rather than a candidate gene is used [2, 6]. Schaid & Sommer [10] suggest the use of two genotype RRs to distinguish between genotypes with one or two copies of allele $a_2$. Valid inference generally requires that the case–control matching be taken into account.

Examination at the level of the individual allele, rather than the genotype, leads to an allelic relative risk. The alleles transmitted and not transmitted by each of the parents are the basis of this risk measure, which is closely related to the transmission/disequilibrium tests (TDT) noted in the following section. However, inference can be complicated by a lack of independence between the parental transmissions, and models that condition on the parental genotypes are more suitable in this case [9].

For a multiallele locus with parents having alleles $a_1a_2$ and $a_3a_4$, there are four possible genotypes that could be observed in their offspring. If the affected child (D) received alleles $a_1$ and $a_3$, then the genotypes $a_2a_4$, $a_1a_4$, and $a_2a_3$ can be taken as control (D) genotypes. Self et al. [11] formulated a general likelihood for a series of independent affected children based on indicator variables for the presence or absence of a given allele in the case and control genotypes. Their approach assumes uniform segregation of gametes apart from the genotype effect, but does not require the parental genotypes to be independent. The likelihood they develop has the same form as that for a logistic regression analysis of a matched case–control study with a single case and three controls.

Features of other family-based case–control designs that include siblings and cousins as controls have been of recent interest [3, 14].

Risk Ratios for Genetic Markers

Genotype, allelic, and haplotype RRs analogous to those described for candidate genes can be estimated using genetic markers. When specific genetic loci for a disease are unknown, finely spaced genetic markers with known locations may be used to screen the whole genome or selected genomic regions to help localize a disease-susceptibility gene.

Allelic information for a single locus can be extended to the multilocus setting by considering haplotypes. When two or more neighboring loci are considered together, a haplotype can be defined as a multilocus analog of an allele. A pattern of alleles from each of several loci that are transmitted together from one parent constitute one haplotype. For two loci, with $n_1$ and $n_2$ alleles, respectively, occurring in a population, there are $n_1 \times n_2$ possible haplotypes that can be expressed as $G = a_ib_j$, where $a_i$ and $b_j$ represent allele types $i, i = 1, \ldots, n_1$, and $j, j = 1, \ldots, n_2$, from the two loci comprising the haplotype. A pair of haplotypes, one inherited from each parent, constitutes a multilocus genotype. A multilocus haplotype of two genetic markers defined in this way can be used to construct a haplotype relative risk analogously to a single genetic marker.

A multilocus haplotype can also be constructed from a genetic marker and disease gene, but the latter is usually unobserved. When two loci forming a haplotype on a parental chromosome are close together, they are less likely to be separated by recombination when a gamete is formed during meiosis. This phenomenon can be exploited in disease–gene localization studies through the modeling of a haplotype consisting of an allele at a known marker location and an unobserved disease allele. Association between particular alleles of a genetic marker and specific alleles...
of the unobserved susceptibility gene that occurs across families in a population is known as allelic association or linkage disequilibrium, whereas the term linkage refers to association that occurs within a family. The absence of tight linkage disequilibrium between alleles at an unobserved causal disease gene locus and measured alleles at marker loci can induce attenuation bias in the RR estimates based on marker alleles [4, 6].

The class of methods known as TDT used in family-based designs involve hypothesis tests that detect linkage only in the presence of allelic association [6, 12]. These methods can also test for association in the presence of linkage, provided correlation between parental transmissions or among related individuals induced by the presence of linkage are taken into account [1, 12].

References


