Am. J. Hum. Genet. 72:636–649, 2003

Improving the Prediction of Complex Diseases by Testing for Multiple Disease-Susceptibility Genes

Quanhe Yang,[1] Muin J. Khoury,[2] Lorenzo Botto,[1] J. M. Friedman,[4] and W. Dana Flanders[3]

[1]National Center on Birth Defects and Developmental Disabilities and [2]Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, and [3]Department of Epidemiology, School of Public Health, Emory University, Atlanta; and [4]Department of Medical Genetics, University of British Columbia, Vancouver

Studies have argued that genetic testing will provide limited information for predicting the probability of common diseases, because of the incomplete penetrance of genotypes and the low magnitude of associated risks for the general population. Such studies, however, have usually examined the effect of one gene at time. We argue that disease prediction for common multifactorial diseases is greatly improved by considering multiple predisposing genetic and environmental factors concurrently, provided that the model correctly reflects the underlying disease etiology. We show how likelihood ratios can be used to combine information from several genetic tests to compute the probability of developing a multifactorial disease. To show how concurrent use of multiple genetic tests improves the prediction of a multifactorial disease, we compute likelihood ratios by logistic regression with simulated case-control data for a hypothetical disease influenced by multiple genetic and environmental risk factors. As a practical example, we also apply this approach to venous thrombosis, a multifactorial disease influenced by multiple genetic and nongenetic risk factors. Under reasonable conditions, the concurrent use of multiple genetic tests markedly improves prediction of disease. For example, the concurrent use of a panel of three genetic tests (factor V Leiden, prothrombin variant G20210A, and protein C deficiency) increases the positive predictive value of testing for venous thrombosis at least eightfold. Multiplex genetic testing has the potential to improve the clinical validity of predictive testing for common multifactorial diseases.

Introduction

The rapid pace of genetic discoveries has resulted in genetic tests for many diseases. A key question is whether genetic tests will be able to predict a healthy person's probability of developing a disease, particularly one of the many common diseases of presumed multifactorial origin. Some researchers suggest that genetic testing will be widely used for this purpose in the near future (Bell 1998; Beaudet 1999; Collins 1999; Evans et al. 2001). Others argue that genetic testing for common diseases will not be useful in practice, because of the incomplete penetrance and low magnitude of risks associated with various genotypes in the population (Holtzman and Marteau 2000; Vineis et al. 2001).

The latter argument is a useful counterbalance to the unrealistic expectation that a single genetic test for, say, cancer or coronary artery disease will revolutionize medicine. However, we believe that this position overstates the intrinsic limitations of genetic testing. The pitfall of such an argument is that it restricts its scope to tests that examine a single genetic factor, whereas simultaneous testing of multiple predisposing alleles is likely to be the standard for multifactorial diseases (Beaudet 1999; Evans et al. 2001). In this article, we show that, if several factors (e.g., genetic loci) play a role in disease etiology, then, under many conditions, evaluating such factors concurrently (e.g., through use of a panel of genetic tests) substantially increases the predictive value for the disease.

A similar result was reported in a recent theoretic examination of populations, using simple additive (multifactorial) models (Pharoah et al. 2002). However, although that finding is of interest, it does not directly apply to the testing of individual patients. Our approach examines the practical use of a test panel of genetic variants with known population frequencies and disease associations to estimate the probability that a healthy person will develop the disease. We describe a general method to generate such probabilities, expand it to include the effect of environmental factors and interactions, and show how the approach performs using plausible simulated data as well as real data for venous thrombosis, a common multifactorial disease.

Received October 9, 2002; accepted for publication December 5, 2002; electronically published February 14, 2003.

Address for correspondence and reprints: Dr. Quanhe Yang, National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, 4770 Buford Highway, Mailstop F-45, Atlanta, GA 30341. E-mail: qyang@cdc.gov

0002-9297/2003/7203-0014$15.00

Methods

We use the likelihood ratio to estimate the posterior probability of disease that is influenced by many factors. The likelihood ratio reflects the probability that a patient with the disease has an observed test result, compared with the probability that a patient without the disease has the same result (Sackett 1991). The likelihood ratio is useful for modeling the contribution of multiple genetic and environmental factors, including interaction effects. In subsequent sections, we describe the main results that will be used for calculating the probability of disease.

Likelihood Ratio

For simplicity, we assume that we are dealing with multiple disease-susceptibility genes, each of which has two alleles. A panel of tests will generate a result for each person, which can be described succinctly by G (g_1, g_2, g_3, … g_n), the vector of test results for the n disease-susceptibility genes (g_1–g_n). If $g_i = 1$ for a positive genetic test result and $g_i = 0$ for a negative result, then each person who is tested can be associated with a string (of length n) of 0s and 1s. For a panel of n tests, there are 2^n theoretical combinations of test results (and 2^n subgroups, each with a different combination of test results, in the population).

If D represents the diseased population and $\overline{D}$ the nondiseased population, one can define the likelihood ratio for any observed value of G as

$$LR(G) = \frac{P(G|D)}{P(G|\overline{D})} , \qquad (1)$$

where $P(G|D)$ represents the probability of G given the presence of disease D and $P(G|\overline{D})$ is the probability of G given the absence of disease D. The likelihood ratio will be higher for combinations of test results that more clearly distinguish people with the disease from those without the disease, thus justifying its frequent use for clinical screening and diagnostic testing (Sackett 1991; Wald and Leck 2000).

Typically, the likelihood ratio in equation (1) has been used to evaluate diagnostic tests by assessing the probability that a disease is present in people with a positive test result (Sackett 1991). We show how the likelihood ratio also can be used to identify people at high risk of *developing* a disease. Such information is useful in prevention activities targeting people who are most likely to develop a disease.

Calculating this likelihood ratio requires recognition of the fact that genetic tests used to predict multifactorial disease are not diagnostic tests. High-risk alleles at any single locus often occur in persons in whom the disease will never develop, and low-risk alleles often occur in patients in whom the disease develops. According to the mutifactorial model, the disease will develop only in people whose combined burden of genetic and environmental risk factors exceeds a certain threshold. Moreover, this threshold may vary with age. In the illustration in this article, we define a single genetic test indicating increased risk for disease as "allele positive" and a single genetic test indicating a decreased risk for disease as "allele negative."

To grasp the concept of computing likelihood ratios for a panel of tests, one can begin with the simpler situation of a single binary test, moving then to a panel of two tests, then three, etc. For the single binary genetic test G (1 or 0), the associated likelihood ratio, $LR(G)$, takes the values $LR(G = 1)$ or $LR(G = 0)$. $LR(G = 1)$ is defined as the likelihood ratio for an allele-positive test, and $LR(G = 0)$ is the likelihood ratio for an allele-negative test. Appendix A shows calculation of the likelihood ratio and other related measures applied in this context.

As mentioned above, for $G = n(g_1, g_2, g_3, \ldots g_n)$ genetic tests, there are 2^n combinations of test results in the population. For example, a panel of two binary genetic tests could have four possible results, and a likelihood ratio can be calculated for each: $LR(g_1 = 0, g_2 = 0)$, $LR(g_1 = 0, g_2 = 1)$, $LR(g_1 = 1, g_2 = 0)$, and $LR(g_1 = 1, g_2 = 1)$. If the n genetic tests (g_1, g_2, g_3, … g_n), are independent, then the joint probability of a given result is the product of the individual probabilities, $P(G|D) = P(g_1|D)P(g_2|D) \ldots P(g_n|D)$. The same is true for $P(G|\overline{D})$. It follows immediately that

$$LR(G) = LR(g_1)LR(g_2) \ldots LR(g_n) , \qquad (2)$$

where

$$LR(g_i) = \frac{P(g_i|D)}{P(g_i|\overline{D})}, \ (i = 1,2, \ldots n) .$$

Thus, the likelihood ratio for a panel of independent tests is simply the product of the likelihood ratios of the individual test results.

When the n genetic tests are not independent, the LR can still be computed, since, by the rule of conditional probability,

$$P(G|D) = P(g_1|g_2, \ldots, g_n, D)P(g_2|g_3, \ldots, g_n, D) \ldots P(g_{n-1}|g_n, D)P(g_n|D) .$$

$P(G|\overline{D})$ can be calculated in an analogous fashion. The expression for the likelihood ratio for multiple genetic

tests that are dependent is more complex but still estimable:

$$LR(G) = LR(g_1|g_2,\ldots,g_n)LR(g_2|g_3,\ldots,g_n)\ldots LR(g_n) \ . \quad (3)$$

When several independent genetic tests for a particular disease are available, one can obtain a combined likelihood ratio through use of equation (2). When several possibly dependent genetic tests exist, one has to use equation (3) and calculate the conditional probabilities in order to get a valid combined likelihood ratio.

Likelihood-Ratio Estimation from Logistic Regression

For a binary disease outcome ($D = 0,1$), assuming a logistic model in the population, we can use logistic regression to calculate the likelihood ratio from a case-control study conducted in the population:

$$\ln LR(G) = \ln\left(\frac{N_{CO}}{N_{CA}}\right) + \alpha_{CC} + \beta G^T = \alpha^* + \beta G^T \ , \quad (4)$$

where α_{CC} and β are the intercept term and logistic regression coefficient of the odds of developing the disease, respectively; N_{CA} is the number of case subjects in the study sample, N_{CO} is the number of control subjects in the study sample, and $\alpha^* = \alpha_{CC} + \ln(N_{CO}/N_{CA})$. To estimate $LR(G)$ using logistic regression in a case-control study, one needs to use the adjusted intercept term, α^*. Appendix B provides a proof of this use of logistic regression to calculate the likelihood ratio from a case-control study. Although we use logistic regression to estimate the likelihood ratio, one could use other link functions (e.g., log linear) instead.

Likelihood Ratio with Covariates and Interaction

So far, we have assumed that each gene independently contributes to the disease and that the population is homogeneous with respect to test results—that is, the probability of having an allele-positive or allele-negative result is the same for every individual. However, such assumptions may not hold. For example, many common diseases are age dependent, and the effect of a certain combination of alleles (and therefore the probability of disease associated with a particular set of test results) may differ depending on exposure to environmental or behavioral factors. In addition, an individual with a strong family history for a particular disease may be more likely to develop that disease than another individual who has the same combination of test results but no family history. Interactions among genetic variants at different loci may also cause dependencies in the results of the test panel.

In this situation, one can estimate the likelihood ratio while adjusting for covariates and including interaction effects. This approach leads to a model with the general form

$$\ln LR(G) = \ln\left(\frac{N_{CO}}{N_{CA}}\right) + \alpha^* + \beta G^T + \gamma X^T + \delta W^T \ , \quad (5)$$

where X is a vector of covariates and W represents interaction effects of multiple binary genetic tests. Failure to consider the effects of some covariates—for example, age as a covariate for an age-dependent disease—may result in a biased estimate of the likelihood ratio.

The variance of the likelihood ratio can be calculated by using the standard delta method based on a Taylor series expansion (see appendix B). The $100(1-\alpha)\%$ CI of the likelihood ratio can be calculated by

$$\exp\left\{\ln LR(G) \pm Z_{1-\alpha/2}\sqrt{\mathrm{Var}[LR(G)]}\right\} \ ,$$

where $Z_{1-\alpha/2}$ is the normal deviate that cuts off appropriate areas in the tails of the standard normal distribution.

Positive and Negative Predictive Value (Posterior Probability)

When using a genetic test to predict the development of a multifactorial disease, we are interested in knowing the probability that the disease will develop in people with an allele-positive result, or $P(D|G)$, and the probability that the disease will not develop in people with an allele-negative result, or $P(\overline{D}\,|G_0)$. $P(D|G)$ is defined as the positive predictive value, or posterior probability, of disease occurrence, and $P(\overline{D}\,|G_0)$ is defined as the negative predictive value. It can be shown that $P(D|G)$ and $P(\overline{D}\,|G_0)$ are functions of the likelihood ratio and of the pretest risk of the disease in the population, $P(D)$:

$$P(D|G) = \frac{LR(G)P(D)}{[1-P(D)]+LR(G)P(D)} \ . \quad (6)$$

Similarly, the negative predictive value can be expressed as

$$P(\overline{D}\,|G_0) = \frac{1}{1+LR(G_0)\frac{P(D)}{1-P(D)}} \ ,$$

where $P(D)$ is the pretest risk of disease or the average risk of disease in the population and $LR(G_0)$ is the likelihood ratio of all allele-negative test results (i.e., the likelihood ratio that all G tests (g_1, g_2,… g_n) take the value of 0). Therefore, one can convert the pretest risk of disease, $P(D)$, to a posterior probability of disease

(positive or negative predictive value) through a set of estimated likelihood ratios from a case-control study. Here we use "positive and negative predictive value" and "posterior probability" interchangeably.

Simulated Data

Using a simulation study, we now illustrate how likelihood ratios can generate the probability of developing disease, on the basis of results from a panel that tests for disease-susceptibility alleles. We simulated a population of one million people and a multifactorial disease with a background risk of 5% (the order of magnitude of common multifactorial diseases such as diabetes or depression). We assume that the risk for developing the disease is influenced by five biallelic disease-susceptibility loci (g_1, g_2, g_3, g_4, and g_5) and one dichotomous environmental exposure, with expected relative risks for the disease of 1.5, 2.0, 2.5, 3.0, 3.5, and 2.0, respectively. We assume that these gene variants and the environmental exposure are all common in the population: 25% for g_1, 20% for g_2, 15% for g_3, 10% for g_4, 5% for g_5, and 15% for the environmental factor. We also assume that the environmental exposure and g_1 interact multiplicatively. Such high frequencies, low associated relative risks, and interaction effects were chosen as plausible scenarios for many multifactorial conditions. We randomly selected a sample of 500 case subjects and 500 control subjects from the population. Choosing a 1:1 case-control ratio is not necessary but simplifies the estimation of likelihood ratio from equation (4) because $\ln(N_{CO}/N_{CA}) = 0$, so that $\ln LR(G) = \alpha + \beta G^T$.

Nomogram

We use the nomogram (fig. 1) to illustrate the increased ability to predict a multifactorial disease, using a panel of genetic tests under a range of scenarios. The nomogram converts the background risk of disease (pretest risk of disease, $P[D]$) to a predicted value (posterior probability of disease occurring, $P[D|G]$), using different values of the likelihood ratio $LR(G)$ (Fagan 1975).

Results

For a multifactorial disease with moderate effects of any single locus (relative risk = 1.5–3.5), any single allele-positive test has limited ability to predict development of the disease (table 1). For example, the likelihood ratio for the genetic test for g_1 alone was computed as $\ln LR(g_1) = \alpha + \beta g_1 = -0.2428 + g_1 \times .7825 = .5397$. The likelihood ratio for an individual who is allele positive for g_1 is given by $\exp(.5397) = 1.72$. For a disease with an overall risk of 5% in the population, the probability of developing the disease, $P(D|G)$, among people with an allele-positive test result for g_1 is

$$P(D|G) = \frac{P(D)LR(g_1)}{1 - P(D) + P(D)LR(g_1)} = \frac{0.05 \times 1.72}{0.95 + 0.05 \times 1.72} = 8.3\% .$$

The variance of the likelihood ratio was calculated using equation (B6) from the covariance matrix generated by the logistic regression analysis as $\mathrm{Var}[LR(g_1)] = 0.00591 + 0.01955 - 0.01182 = 0.01364$.

Similarly, we can estimate the likelihood ratio for the g_1 allele-negative test result as $\exp(-0.2428) = 0.784$ and the corresponding probability of *not* developing the disease among people with an allele-negative result for g_1, $P(\overline{D}|G_0) = 1/1.0413 = 96.0\%$.

The simulated data in table 1 show how a panel of genetic tests improves the positive predictive value under increasingly inclusive scenarios (i.e., testing g_1 only; combined testing of g_1 and g_2; combined testing of g_1, g_2, and g_3; and so on). Figure 1 displays these results on a nomogram that can also be used to take into account the effect of different pretest risks of disease, $P(D)$. The posterior probability of disease increases with the number of informative genetic tests done concurrently, with more than a 10-fold increase between a test for g_1 only (posterior probability 8%) and multiple genetic tests and an environmental risk factor (posterior probability 89%).

For any given test panel, the pretest risk of the disease in the population also has an important impact on the predictive value. For example, if the pretest risk of the disease increases from 5% to 10%, such as may occur when people with a first-degree relative affected with the disease are tested, the posterior probability would increase from 8% to 16% for a single genetic test and from 89% to ~94% for the full panel of tests for five genes and one environmental exposure (fig. 2).

Venous Thrombosis: An Example Using Real Data

In a review article, Seligsohn and Lubetsky (2001) discussed genetic predisposition to venous thrombosis and proposed a set of tests for inherited thrombophilia. Most inherited thrombophilia can be attributed to either failure to control the generation of thrombin or impaired neutralization of thrombin. Factor V Leiden (the Arg506Gln substitution in factor V), the G20210A variant of prothrombin (the G20210A mutation in the 3′ UTR of the prothrombin gene), and deficiencies of proteins C or S are associated with decreased control of

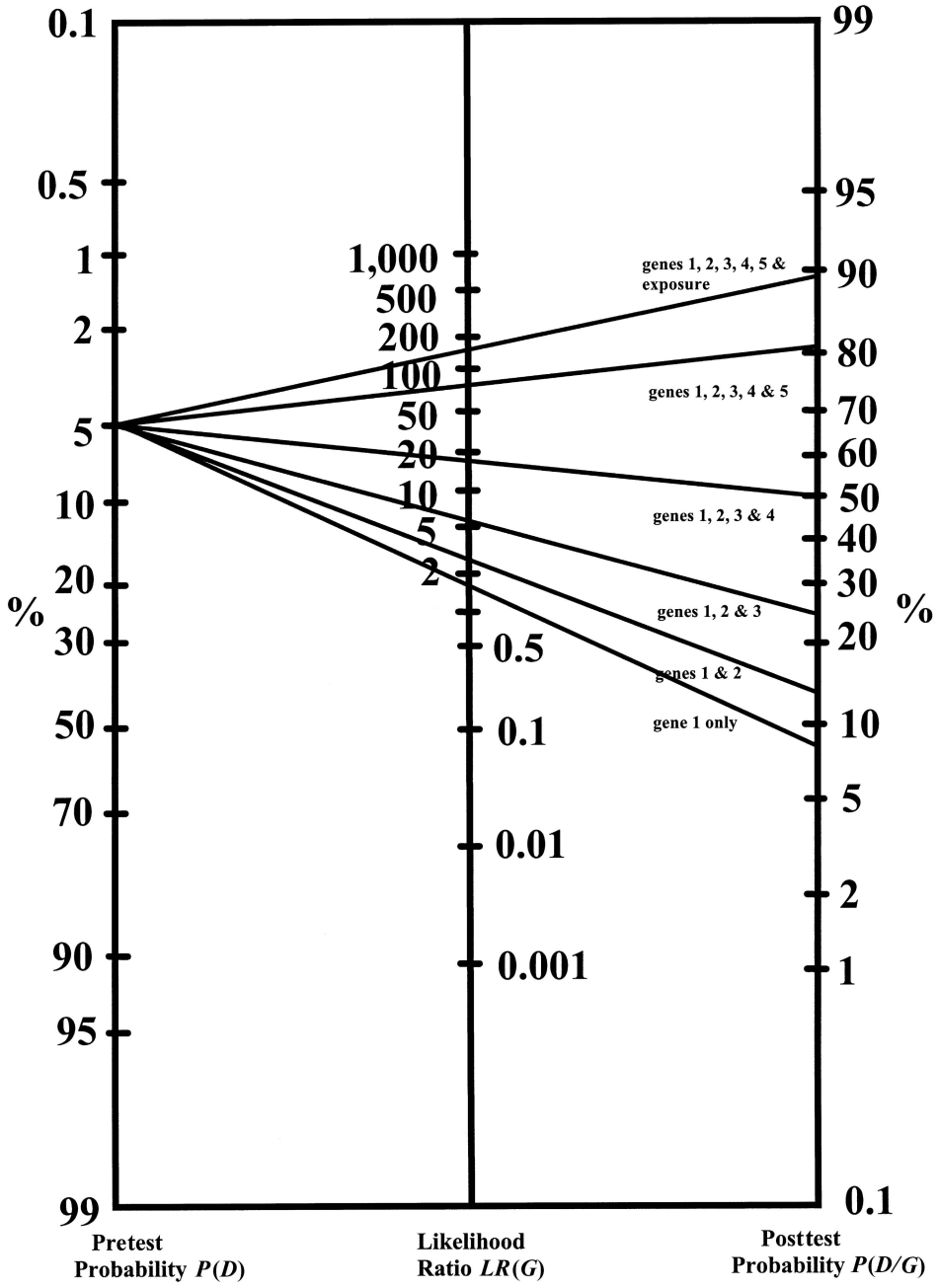

Figure 1 Power of a panel of genetic tests and exposure on predictability of the common disease (simulated data)

thrombin generation. Deficiency of antithrombin leads to decreased neutralization of thrombin. Seligsohn and Lubetsky (2001) pooled 30 studies of genetic susceptibility to venous thrombosis and presented data on the frequency of various inherited thrombophilias among healthy subjects and groups of patients with venous thrombosis.

To demonstrate the likelihood-ratio approach to predicting the probability of disease development, we first derive the relevant allele frequencies for factor V Leiden, the G20210A prothrombin gene variant, and protein C deficiency among patients with venous thrombosis, using data from Seligsohn and Lubetsky's (2001) review (table 2). For factor V Leiden and the G20210A prothrombin gene variant, we included only white subjects, because of the very low frequency of these variants among Asians

Table 1

Likelihood Ratios, 95% CIs of Likelihood Ratios, and Posterior Probability of Developing Disease for Single and Multiple Genetic Susceptibility Tests and an Environmental Exposure

Environmental Exposure/Genetic Tests and Coefficient[a]	Estimate	SE	*P*	Expected Relative Risk of Genotype and Environmental Exposure	Expected Gene Frequency and Environmental Exposure Prevalence in Population (%)	*LR* (95% CI)	Posterior Probability of Developing Disease[b] (%)
One-gene test:							
α	−.2428	.0769	.002			1.72 (1.37–2.16)	8.3
g_1	.7825	.1398	<.0001	1.5	25		
α	−.1237	.0712	.082			1.61 (1.22–2.13)	7.8
g_2	.6015	.1591	<.0002	2.0	20		
α	−.1652	.0704	.019			2.10 (1.55–2.85)	9.9
g_3	.9062	.1708	<.0001	2.5	15		
α	−.1456	.0681	.033			2.75 (1.88–4.03)	12.6
g_4	1.1572	.2062	<.0001	3.0	10		
α	−.1070	.0661	.106			3.72 (2.21–6.26)	16.4
g_5	1.4213	.2736	<.0001	3.5	5		
Two-gene tests (selected examples):							
α	−.3879	.0855	<.0001				
g_1	.8151	.1413	<.0001	1.5	25	2.95 (2.07–4.20)	13.4
g_2	.6543	.1620	<.0001	2.0	20		
α	−.4079	.0839	<.0001				
g_1	.7824	.1418	<.0001	1.5	25	3.60 (2.49–5.21)	15.9
g_3	.9062	.1732	<.0001	2.5	15		
α	−.2795	.0777	<.0003				
g_2	.5741	.1614	<.0004	2.0	20	3.26 (2.20–4.84)	14.6
g_3	.8868	.1718	<.0001	2.5	15		
Three-gene tests (selected examples):							
α	−.5455	.0921	<.0001				
g_1	.8174	.1434	<.0001	1.5	25	6.02 (3.79–9.56)	24.1
g_2	.6333	.1646	<.0001	2.0	20		
g_3	.8898	.1748	<.0001	2.5	15		
g_1	.8021	.1416	<.0001	1.5	25		
g_2	.6904	.1677	<.0001	2.0	20	19.0 (10.5–35.6)	50.0
g_3	.9407	.1775	<.0001	2.5	15		
g_4	1.2287	.2134	<.0001	3.0	10		
Five-gene tests:							
α	−.8405	.1031	<.0001				
g_1	.8502	.1488	<.0001	1.5	25		
g_2	.6629	.1709	.0001	2.0	20		
g_3	.9501	.1798	<.0001	2.5	15	77.6 (33.2–181.2)	80.3
g_4	1.2396	.2160	<.0001	3.0	10		
g_5	1.4886	.2853	<.0001	3.5	5		
Genes and exposure:							
α	−1.0563	.1121	<.0001				
g_1	.6866	.1539	<.0001	1.5	25		
g_2	.7166	.1733	<.0001	2.0	20		
g_3	.9593	.1829	<.0001	2.5	15	151.7 (64.9–354.8)	88.9
g_4	1.2103	.2205	<.0001	3.0	10		
g_5	1.5513	.2888	<.0001	3.5	5		
e	.9539	.1616	<.0001	2.0	15		

[a] α = Intercept of logistic regression; g_1–g_5 = coefficients of logistic regression for genetic tests for genes 1–5, respectively; e = coefficient of logistic regression for a dichotomous environmental exposure variable.

[b] Values were obtained from equation (6), using simulated data with a 5.0% background risk for the disease.

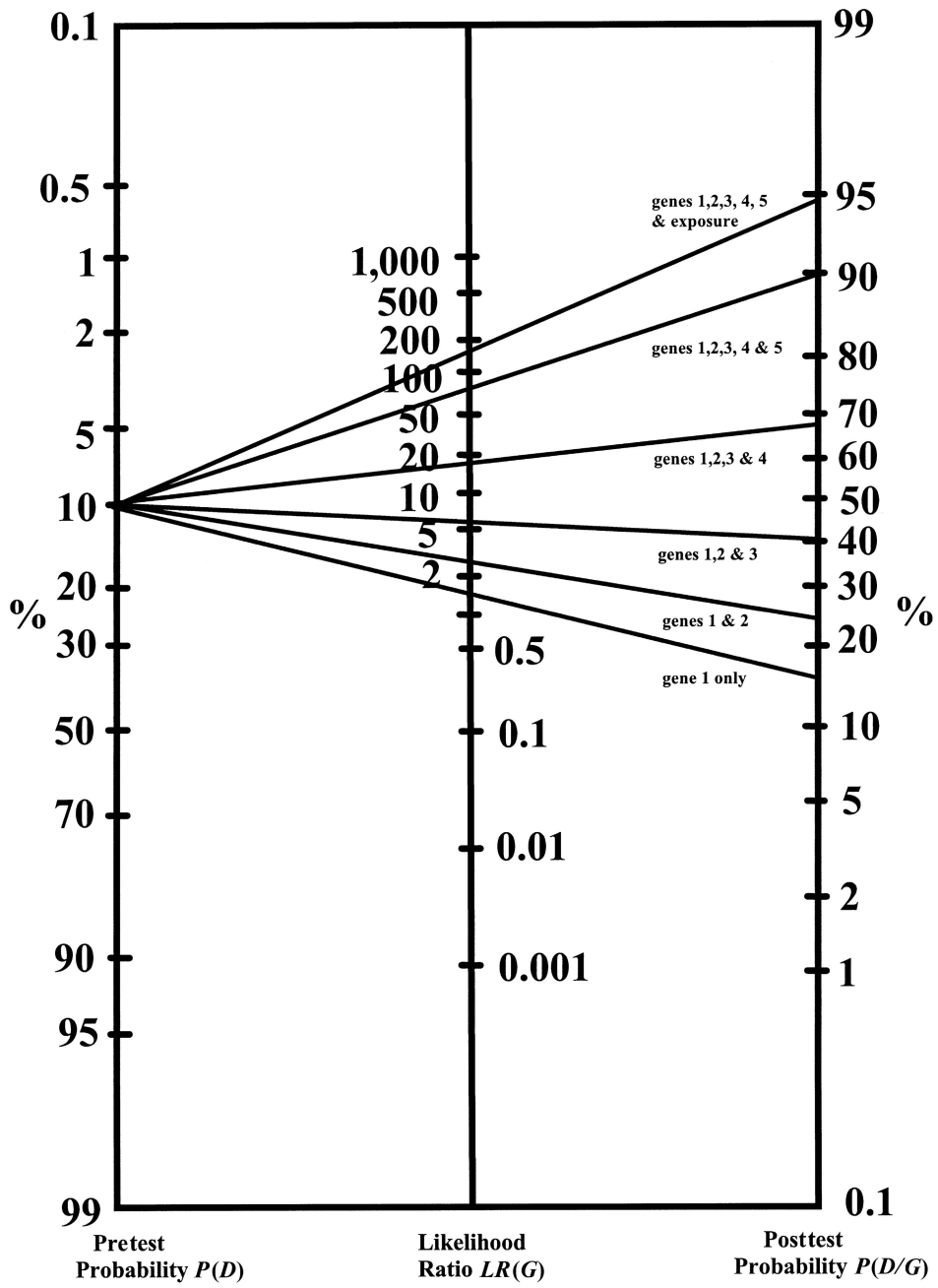

Figure 2 Impact of prevalence of disease on posterior probability of disease (simulated data)

and Africans. In this illustration, we treat these meta-analysis results as a valid estimate of the risk odds ratio, as would be derived from a well-designed case-control study. Appendix B provides a proof that a valid estimate of the likelihood ratio can indeed be obtained from an well-designed case-control study. We calculated unadjusted likelihood ratios for each test through use of logistic regression, assuming an independent effect of each allele tested. We then converted these results to the posterior probability of developing disease, by the method described above.

The computation of likelihood ratios using logistic regression (table 2) is straightforward. For example, the likelihood ratio for the allele-positive test for factor V Leiden among healthy subjects and unselected patients with venous thrombosis is obtained by $\ln(LR) = \ln(N_{CO}/N_{CA}) + \alpha + \beta = \ln(16{,}150/1{,}142) - 2.809 + 1.526 = 1.3669$, where α and β are intercept term and

Table 2

Distribution of Inherited Thrombophilia among Healthy Subjects and Unselected and Selected Patients with Venous Thrombosis, by Status of Factor V Leiden, the G20210A Prothrombin Gene Mutation, and Protein C Deficiency

Inherited Thrombophilia	No. of Healthy Subjects	No. of Patients	*LR* (95% CI)	Posterior Probability of Developing Disease (%)
Panel A (healthy subjects and unselected patients):				
Factor V Leiden:				
+	775	215	3.9 (3.4–4.6)	.8
−	15,375	927		
G20210A prothrombin gene mutation:				
+	11,610	205	2.6 (2.2–3.1)	.5
−	322	2,679		
Protein C deficiency				
+	45	74	12.3 (8.5–17.8)	2.4
−	15,025	1,937		
Combined tests[a]			126.8	20.3
Panel B (healthy subjects and selected patients)[b]:				
Factor V Leiden:				
+	775	65	8.4 (6.5–10.8)	1.6
−	15,375	97		
G20210A prothrombin gene mutation:				
+	11,610	88	5.9 (4.7–7.5)	1.2
−	322	463		
Protein C deficiency:				
+	45	37	16.2 (10.5–25.0)	3.1
−	15,025	730		
Combined tests[a]			801.8	61.6

Note.—Derived from Seligsohn and Lubetsky (2001).

[a] Likelihood ratios for combined tests were estimated by assuming independence of the tests.

[b] Patients who met the following criteria were selected: age <50 years, family history of venous thrombosis, personal history of recurrent thrombotic events, and absence of acquired risk factors except for pregnancy or the use of oral contraceptives. The estimates of likelihood ratio from the selected patients are likely to be biased because the case subjects (selected patients) are not comparable to the noncase subjects (healthy subjects).

estimated coefficient of logistic regression. The likelihood ratio is calculated by exponentiating this result, $LR = \exp(1.3669) = 3.9$. The variance of the likelihood ratio is $\mathrm{Var}(LR) = 0.001144 + 0.007085 - 0.00228 = 0.005949$, and the 95% CI of the likelihood ratio is $\exp(1.3669 \pm 1.96\sqrt{0.005949}) = (3.37, 4.56)$.

To estimate the posterior probability for venous thrombosis (i.e., the probability of developing the disease, $P[D|G]$) using the likelihood ratio, one must know the pretest risk of the disease in the general population. We recognize that the risk for venous thrombosis varies with age (Ridker et al. 1997) and that it is preferable to include age as a covariate in the model, estimate age-specific likelihood ratios, and convert these likelihood ratios to age-specific posterior probabilities of disease. However, many studies have estimated the overall incidence of venous thrombosis to be 1.5–2 per 1,000 person-years in the general population (Nordstrom et al. 1992; Hansson et al. 1997; White et al. 1998), and, to simplify this demonstration, we assume that the pretest risk of venous thrombosis is 2 per 1,000. We also assume that the effect of each susceptibility gene is independent and that all interactive effects are purely multiplicative.

Each genetic test provides limited predictive information about the probability of developing venous thrombosis. The posterior probabilities of disease range from 0.5% to 3.1% for each test alone. However, the posterior probabilities of venous thrombosis occurring increases to 20.3% when estimated with unselected patients and to 61.6% when estimated with selected patients, an increase of >8-fold for unselected patients and >20-fold for selected patients.

Discussion

We have shown that using a panel of genetic tests can substantially improve the ability to predict the risk of developing a multifactorial disease, compared with using just one test, providing that the panel includes factors

that contribute to the disease. The argument is still valid if the assessment includes not only testing for susceptibility alleles but also information on environmental exposures or other predisposing factors. One can use likelihood ratios to integrate such genetic and environmental assessments into a summary estimate of the risk that a particular healthy person will develop the disease.

Combining information from multiple risk factors to predict the probability of disease development is not new. For example, Gail et al. (1989) used a proportional hazards model to estimate individual probabilities of developing breast cancer, on the basis of factors such as age at menarche, age at first live birth, number of previous breast biopsies, and number of first-degree relatives with breast cancer. Estimating likelihood ratios through use of logistic regression with covariates has been proposed in clinical diagnostic tests (Coughlin et al. 1992; Simel et al. 1993). A method similar to the one we propose is used routinely in pregnancy screening, to estimate the risk of fetal Down syndrome, on the basis of multiple maternal serum markers, maternal age, and other factors (Wald and Leck 2000).

We show that an individual patient's risk of developing a multifactorial disease can be calculated from case-control data by means of likelihood ratios estimated using logistic regression. This approach permits the simultaneous use of information from many different genetic tests as well as from environmental risk factors, age, personal medical history, and family history. When all such information is taken into account, the estimated likelihood ratio can easily be converted to the posterior probability of developing the disease.

The nomogram graphically illustrates the conditions that improve prediction—namely, increasing the number of risk factors that are considered and focusing on groups with a higher background risk for the disease. For a common disease (affecting $\geqslant$10% of the population), a positive test panel associated with a combined likelihood ratio of 81 would strongly predict the probability of developing the disease (in excess of 90% posterior probability). A likelihood ratio of this magnitude can be achieved by using a small panel of disease-susceptibility alleles with moderate effects or by using fewer alleles with relatively strong effects. The availability of multiplex genetic testing by efficient automated methods (Southern 1996; Pennisi 1999) offers the prospect of assessing dozens or hundreds of alleles simultaneously and thereby identifying individuals at very high risk of developing a particular disease, even if the contribution of each gene to the risk is small. Focusing this testing on higher-risk groups, such as people with a positive family history, can increase the prediction probabilities even further because of the higher a priori risk for a disease among people with a positive family history. For a group whose a priori risk of developing a disease is 15% instead of 10%, for example, a combined likelihood ratio of 51 (instead of 81) would be sufficient to reach the same 90% posterior probability of disease development.

Although our findings indicate that prediction probability improves when common diseases are examined, considering multiple genetic risk factors simultaneously also improves the prediction probability for rarer conditions. For example, in venous thrombosis, which is relatively uncommon (1.5–2 per 1,000 in the population), an appropriately tailored panel of genetic tests combined with age and other potential risk factors could achieve a positive predictive value in excess of 90%.

These findings lead us to two considerations. First, methods based on likelihood ratios can be useful and effective tools to evaluate the probability of developing a disease in relation to multiple genetic and environmental factors and their interactions. Second, the ability of genetic tests to predict multifactorial diseases is not inherently low but depends on how many factors are considered and the characteristics of each factor with respect to population frequency, associated risks, and interactions. As knowledge of these factors and their associated parameters improves, so will the ability to predict the probability of developing diseases. At that point, the major limiting factor in prediction might be the background risk in the population (the disease incidence), so that, contrary to some views, common multifactorial diseases might be more reliably predictable than conditions that are neither common nor multifactorial. This view is supported by the fact that a positive panel of tests for common alleles and relatively weak risk factors, when taken as a whole, may be as informative as testing positive for a single, strong risk factor.

Such considerations are valid to the extent that the model implicit in the test panel correctly reflects the underlying etiology of the disease. Thus, valid prediction is predicated upon correctly including relevant gene variants in the panel and valid exposures in the global assessment, as well as upon correctly defining the dependencies (e.g., interactions) among gene variants and environmental factors.

In our illustration, we described and simulated scenarios in which all gene variants conferred an increased risk for disease. The simultaneous presence of genotypes that confer a lower risk adds complexity to the scenario but can easily be included in the calculation.

An important consideration in testing for multiple weak genetic predispositions is the trade-off between precision of the prediction and the size of the group of people to whom the prediction applies. The number of people identified as being at highest risk decreases as the precision of prediction increases and, generally, as the number of (independent) component tests increases. A panel

of n tests, for example, can generate up to 2^n combinations of test results, and their distribution in the population depends on the population frequencies of the genes in the panel. When independent loci are assumed, the proportion of the population with a given combination of test results is equal to the product of the relative frequency of each component result (allele). Testing is most predictive for people who carry all of the susceptibility alleles tested by the panel, but these people will probably represent only a small proportion of those who eventually develop the disease. For others, the ability to predict disease will decrease with the number of "at-risk" alleles carried. With the rapid advancement of genomic technology, a large number of genetic tests will likely become available for some multifactorial diseases. As the number of genetic tests increases, application of the likelihood-ratio approach to many different combinations of allele-positive and allele-negative results would generate a more or less continuous distribution of posterior probabilities of disease. Most cases of the disease would occur among people who are at high risk (as measured by the posterior probability) (Pharoah et al. 2002). The decision about the appropriate cutoff point for public health interventions or individual risk factor modifications is a complex one and will likely depend on the nature of the disease (e.g., its mortality and morbidity), the effectiveness and cost of treatment, and the cost-effectiveness of screening (Bell 1998; Evans et al. 2001; Guttmacher and Collins 2002).

The use of logistic regression to estimate likelihood ratios permits investigators to include important covariates in the model. Age, sex, personal medical history, and family history frequently influence an individual's risk of developing a multifactorial disease. Failure to take these covariates into account may result in biased estimates of the likelihood ratios and inaccurate calculation of the posterior probability of disease. The logistic regression method we propose for estimating likelihood ratios assumes a multiplicative relationship of the risk factors. Additive effects of different genes can also be considered in this model, with an alternative parameterization of gene-gene and gene-environment interaction (Hosmer and Lemeshow 1992; Botto and Khoury 2001). One of the shortcomings of a multiplicative model is that unrealistically high risk estimates may be obtained when many factors are considered simultaneously. The investigators must be cautious when specifying their models with multiple genetic and risk factors and when interpreting results from such models.

Our models assume a homogeneous population with fixed frequencies of alleles and background (pretest) disease risk. In fact, a population may actually be composed of subpopulations, such as racial groups, with different allele frequencies and background risks of disease. Under these circumstances, a stratified analysis can be used to generate valid estimates of likelihood ratios by logistic regression.

We have focused on the clinical validity of genetic tests in this study. Clinical validity, which measures how well an allele-positive result identifies people who will develop the disease and how well an allele-negative result identifies those who will not develop the disease, is an important criterion for safe and effective genetic testing. However, it is important to point out that our estimates of clinical validity depend on the appropriateness of the model, including the multiplicative assumption and the assumption that the genetic and nongenetic factors and their interactions correctly reflect the underlying etiology of the disease. Other important aspects of genetic testing that we did not examine here include analytical validity, clinical utility, and ethical, legal, and social implications (Holtzman et al. 1997, 1998; Barber 1998; Bell 1998).

Finally, we wish to emphasize that using a combination of risk factors (whether genetic or environmental or both) to derive a combined prediction probability requires knowledge of the individual and joint risks. This implies that one knows not only the risk associated with each genotype or environmental exposure but also the strength of each interaction. If such data are lacking, estimates of summary risks would be incomplete and possibly misleading. Unfortunately, however, such data are lacking for most conditions. The clinical and epidemiologic communities can contribute to filling these gaps and improving the prediction of multifactorial diseases by collecting, presenting, and analyzing data on multiple genetic and environmental factors in ways that allow the determination of joint risks and interactions.

Appendix A

Genetic Test Result	No. of People Who Develop the Disease	No. of People Who Do Not Develop the Disease
Allele positive	a	c
Allele negative	$\underline{b}$	$\underline{d}$
Total	N_1	N_2

The positive likelihood ratio ($LR+$) is a ratio between the probability of allele-positive tests among those who develop the disease and the probability of allele-positive tests among those who do not develop disease. It can be calculated as follows:

$$LR+ = \frac{P(G = 1|D)}{P(G = 1|\overline{D})} = \frac{\text{sensitivity}}{1 - \text{specificity}} ,$$

where sensitivity (the probability that people who develop the disease are allele positive) is equal to $P(G = 1|D) = a/N_1$ and specificity (the probability that people who do not develop disease are allele negative) is equal to $P(G = 0|\overline{D}) = d/N_2$.

The negative likelihood ratio ($LR-$) is the ratio between the probability of allele-negative tests among those who develop disease and the probability of allele-negative tests among those who do not develop disease. It can be calculated as follows:

$$LR- = \frac{P(G = 0|D)}{P(G = 0|\overline{D})} = \frac{1 - \text{sensitivity}}{\text{specificity}} .$$

Positive predictive value (PPV) is the probability of developing the disease, given an allele-positive result, and is calculated as

$$\text{PPV} = P(D|G = 1) = \frac{P(G = 1|D)P(D)}{P(G = 1|D)P(D) + P(G = 1|\overline{D})P(\overline{D})} .$$

Negative predictive value (NPV) is the probability of not developing the disease, given an allele-negative result, and is calculated as

$$\text{NPV} = P(\overline{D}|G = 0) = \frac{P(G = 0|\overline{D})P(\overline{D})}{P(G = 0|\overline{D})P(\overline{D}) + P(G = 0|D)P(D)} .$$

We use positive and negative predictive value and posterior probability of developing disease interchangeably.

Appendix B

Estimating the Likelihood Ratio from a Case-Control Study

Assuming dichotomous disease outcome ($D = 0,1$) and a logistic model in the population, we can model the probability of disease for a given panel of genetic tests as (McCullagh and Nelder 1989)

$$\ln\left[\frac{P(D|G)}{P(\overline{D}|G)}\right] = \alpha_{\text{pop}} + \beta G^T . \tag{B1}$$

Applying Bayes's theorem, we have

$$\ln\left[\frac{P(D|G)}{P(\overline{D}|G)}\right] = \ln\left[\frac{P(G|D)P(D)}{P(G|\overline{D})P(\overline{D})}\right] = \ln\left[\frac{P(G|D)}{P(G|\overline{D})}\right] + \ln\left[\frac{P(D)}{P(\overline{D})}\right] .$$

Therefore, the likelihood ratio is

$$\ln LR_{\text{pop}}(G) = \ln\frac{N_{\overline{D}}}{N_D} + \ln\frac{P(D|G)}{P(\overline{D}|G)} = \ln\frac{N_{\overline{D}}}{N_D} + \alpha_{\text{pop}} + \beta G^T (\text{from eq. [B1]}) = \alpha^*_{\text{pop}} + \beta G^T ,$$

where α_{pop} is the intercept term in the population logistic model (background disease risk), N_D is the number of people in the population who develop the disease, $N_{\overline{D}}$ is the number of people in the population who do not develop the disease, $P(D) = N_D|(N_D + N_{\overline{D}})$, and $\alpha^*_{\text{pop}} = \alpha_{\text{pop}} + \ln(N_{\overline{D}}|N_D)$ (Albert 1982).

To prove the validity of estimating likelihood ratio from a case-control study, we introduce the dummy variable S to indicate whether an individual is selected for the case-control sample and denote the sampling fraction as $f_1 = P(S = 1|D)$ and $f_0 = P(S = 1|\overline{D})$. It is essential that the risk odds ratio in the case-control study estimates the risk ratio and the probability of being selected for a sample is independent of genotype in both those with and without the disease—that is, $P(S = 1|D,G) = P(S = 1|D)$ and $P(S = 1|\overline{D},G) = P(S = 1|\overline{D})$. We can compute the probability of disease, given a particular set of genetic test results, using a logistic model for the sample as

$$\ln\left[\frac{P(D|G,S = 1)}{P(\overline{D}|G,S = 1)}\right] = \ln\left[\frac{P(D|G)P(S = 1|D)/P(S|G)}{P(\overline{D}|G)P(S = 1|\overline{D})/P(S|G)}\right] = \ln\left[\frac{P(D|G)}{P(\overline{D}|G)}\right] + \ln\left(\frac{f_1}{f_0}\right) \tag{B2}$$

after cancellation of the denominator. Substitution of equation (B1) into equation (B2) gives

$$\ln\left[\frac{P(D|G,S = 1)}{P(\overline{D}|G,S = 1)}\right] = \alpha_{\text{pop}} + \beta G^T + \ln\left(\frac{f_1}{f_0}\right) = \alpha_{\text{CC}} + \beta G^T , \tag{B3}$$

where $\alpha_{\text{CC}} = \alpha_{\text{pop}} + \ln(f_1/f_0)$. Thus, the logistic model continues to apply in the sample with the same β coefficient but with an adjusted $\alpha^* = \alpha_{\text{pop}} + \ln(f_1/f_0)$ (Breslow et al. 1980).

Similar to the derivation of likelihood ratio estimated using logistic regression in the population, the likelihood ratio in the case-control study population is found to be

$$\ln LR_{\text{CC}}(G) = \ln\frac{N_{CO}}{N_{CA}} + \ln\frac{P(D|G)}{P(\overline{D}|G)} = \ln\frac{N_{CO}}{N_{CA}} + \alpha_{CC} + \beta G^T , \tag{B4}$$

where $\alpha_{\text{cc}} = \alpha_{\text{pop}} + \ln(f_1/f_0)$ is the intercept term estimated from a case-control study, as shown in equation (B3). Because

$$\ln\left(\frac{f_1}{f_0}\right) = \ln\left(\frac{N_{\text{CA}}/N_D}{N_{\text{CO}}/N_{\overline{D}}}\right) = \ln\left(\frac{N_{\overline{D}}}{N_D}\right) - \ln\left(\frac{N_{\text{CO}}}{N_{\text{CA}}}\right) , \tag{B5}$$

substitution of equation (B5) into equation (B4) gives $\ln LR_{\text{pop}}(G) = \ln LR_{\text{cc}}(G)$.

Variance Estimation of the Likelihood Ratio

From equation (5) in the text and the delta method, the variance of the likelihood ratio with covariates and interaction (for dichotomous genetic tests and interaction) is given by (Kleinbaum 1998)

$$\begin{aligned}\mathrm{Var}\,[LR(G)] &= \mathrm{Var}\,(\alpha_{CC} + \beta + \gamma X^T + \delta)\\ &= \mathrm{Var}\,(\alpha_{CC}) + \sum_{i=1}^{n}\mathrm{Var}\,(\beta_i) + \sum_{i=1}^{p_1} X_i^2\,\mathrm{Var}\,(\gamma_i)\\ &\quad + \sum_{i=1}^{p_2}\mathrm{Var}\,(\delta_i) + 2\sum_{i=1}^{n}\mathrm{Cov}\,(\alpha_{CC},\beta_i) + 2\sum_{i=1}^{p_1} X_i\,\mathrm{Cov}\,(\alpha_{CC},\gamma_i) + 2\sum_{i=1}^{p_2}\mathrm{Cov}\,(\alpha_{CC},\delta_i)\\ &\quad + 2\sum_{i=1}^{n}\sum_{i=1}^{p_1} X_i\,\mathrm{Cov}\,(\beta_i,\gamma_i) + 2\sum_{i=1}^{n}\sum_{i=1}^{p_2}\mathrm{Cov}\,(\beta_i,\delta_i) + 2\sum_{i=1}^{p_1}\sum_{i=1}^{p_2} X_i\,\mathrm{Cov}\,(\gamma_i,\delta_i)\\ &\quad + 2\sum_{i<i'}\sum\ \mathrm{Cov}\,(\beta_i,\beta_{i'}) + 2\sum_{i<i'}\sum\ \mathrm{Cov}\,(\gamma_i,\gamma_{i'}) + 2\sum_{i<i'}\sum\ \mathrm{Cov}\,(\delta_i,\delta_{i'})\ , \end{aligned} \tag{B6}$$

where α_{CC} is the intercept of the logistic regression, β is the coefficient of each binary genetic test, X is a vector of covariates, γ is the associated logistic regression coefficient, and δ represents the interaction effects of multiple binary genetic tests. As a general guideline, any two genes in the same pathway would be unlikely to be independent; if they function in different pathways, one would generally assume that they are independent. To evaluate whether inclusion of the interaction terms is necessary, one can use likelihood-ratio tests or Wald χ^2 statistics and associated P values with respect to a χ^2 distribution with 1 df (Greenland 1983). Most statistical analysis programs for logistic regression offer a variance-covariance matrix that can be used to calculate the variance of the likelihood ratio.

References

Albert A (1982) On the use and computation of likelihood ratios in clinical chemistry. Clin Chem 28:1113–1119

Barber JC (1998) Code of practice and guidance on human genetic testing services supplied direct to the public. Advisory Committee on Genetic Testing. J Med Genet 35:443–445

Beaudet AL (1999) 1998 ASHG presidential address: making genomic medicine a reality. Am J Hum Genet 64:1–13

Bell J (1998) The new genetics in clinical practice. BMJ 316: 618–620

Botto LD, Khoury MJ (2001) Commentary: facing the challenge of gene-environment interaction: the two-by-four table and beyond. Am J Epidemiol 153:1016–1020

Breslow NE, Day NE, Davis W, Estève J, International Agency for Research on Cancer (1980) Statistical methods in cancer research. IARC Press, Lyon, France

Collins FS (1999) Shattuck lecture—medical and societal consequences of the Human Genome Project. N Engl J Med 341:28–37

Coughlin SS, Trock B, Criqui MH, Pickle LW, Browner D, Tefft MC (1992) The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. J Clin Epidemiol 45:1–7

Evans JP, Skrzynia C, Burke W (2001) The complexities of predictive genetic testing. BMJ 322:1052–1056

Fagan TJ (1975) Letter: nomogram for Bayes theorem. N Engl J Med 293:257

Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81:1879–1886

Greenland S (1983) Tests for interaction in epidemiologic studies: a review and a study of power. Stat Med 2:243–251

Guttmacher AE, Collins FS (2002) Genomic medicine: a primer. N Engl J Med 347:1512–1520

Hansson PO, Welin L, Tibblin G, Eriksson H (1997) Deep vein thrombosis and pulmonary embolism in the general population. "The Study of Men Born in 1913." Arch Intern Med 157:1665–1670

Holtzman NA, Marteau TM (2000) Will genetics revolutionize medicine? N Engl J Med 343:141–144

Holtzman NA, Murphy PD, Watson MS, Barr PA (1997) Predictive genetic testing: from basic research to clinical practice. Science 278:602–605

Holtzman NA, Watson MS (eds) (1998) Promoting safe and effective genetic testing in the United States: final report of the task force on genetic testing. Johns Hopkins University Press, Baltimore

Hosmer DW, Lemeshow S (1992) Confidence interval estimation of interaction. Epidemiology 3:452–456

Kleinbaum DG (1998) Applied regression analysis and other multivariable methods. Duxbury Press, Pacific Grove, CA

McCullagh P, Nelder JA (1989) Generalized linear models. Chapman and Hall, London

Nordstrom M, Lindblad B, Bergqvist D, Kjellstrom T (1992) A prospective study of the incidence of deep-vein thrombosis

within a defined urban population. J Intern Med 232:155–160

Pennisi E (1999) DNA chips give new view of classic test. Science 283:17–18

Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA (2002) Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet 31:33–36

Ridker PM, Glynn RJ, Miletich JP, Goldhaber SZ, Stampfer MJ, Hennekens CH (1997) Age-specific incidence rates of venous thromboembolism among heterozygous carriers of factor V Leiden mutation. Ann Intern Med 126:528–531

Sackett DL (1991) Clinical epidemiology: a basic science for clinical medicine. Little Brown, Boston

Seligsohn U, Lubetsky A (2001) Genetic susceptibility to venous thrombosis. N Engl J Med 344:1222–1231

Simel DL, Samsa GP, Matchar DB (1993) Likelihood ratios for continuous test results: making the clinicians' job easier or harder? J Clin Epidemiol 46:85–93

Southern EM (1996) DNA chips: analysing sequence by hybridization to oligonucleotides on a large scale. Trends Genet 12:110–115

Vineis P, Schulte P, McMichael AJ (2001) Misconceptions about the use of genetic tests in populations. Lancet 357: 709–712

Wald NJ, Leck I (2000) Antenatal and neonatal screening. Oxford University Press, New York

White RH, Zhou H, Romano PS (1998) Incidence of idiopathic deep venous thrombosis and secondary thromboembolism among ethnic groups in California. Ann Intern Med 128: 737–740