Simulation Studies of Segregation Analysis: Application to Two-Locus Models

DAVID A. GREENBERG[1]

SUMMARY

We tested the power of a segregation analysis method (first proposed by Elandt-Johnson) to distinguish between single-locus and two-locus models, with and without environmentally caused reduced penetrance. We also looked at the effect of ascertainment probability on the analysis and at the proband-conditioned ascertainment correction proposed by Cannings and Thompson. We found that: (1) the segregation analysis has sufficient power to distinguish between the fully-penetrant double-recessive (RR) model and the fully-penetrant single-locus dominant and recessive models; (2) the method can also distinguish fairly well between the dominant-recessive (DR) and RR models, even when one does not take into account the population pıɘvalence; (3) the method has much less power to distinguish between the fully-penetrant RR model and the single-locus models with reduced penetrance; (4) when environmental penetrance is taken account of in the analysis, the power of the method to distinguish between the one- and two-locus models improved substantially; (5) the estimates of ascertainment probability, π, were robust, regardless of the model under which the data were generated; and (6) the Cannings-Thompson approach to ascertainment correction worked well only when the π used to generate the data was less than .1.

INTRODUCTION

Segregation analysis has been used in human genetics for some time, and computer programs, some of them extremely complex, have been written to do the necessary

Received March 16, 1983; revised June 17, 1983.

This work was supported in part by the Veterans Administration, contract N01-NS-02332 from the Department of Health and Human Services, and grant AM-31775 from the U.S. Public Health Service.

[1] Wadsworth VA Medical Center Neurology Service, Room 3405, Los Angeles, CA 90073.

calculations. However, the power of the models and statistical tests incorporated in these programs has seldom been tested because of their complexity and the enormous quantities of computer time required.

We recently [1] used a segregation analysis technique to investigate the inheritance of coeliac disease. This method, originally described by Elandt-Johnson [2], was developed specifically to test whether the inheritance of a disease was due to two epistatic loci. We assumed that there were two alleles (a normal and a disease or trait allele) at each locus and that the inheritance at each locus was either Mendelian dominant or recessive. In the case of coeliac disease, not only is there immunologic evidence for the existence of two loci [3, 4], but it also was possible to eliminate the hypotheses of single-locus dominant and recessive inheritance with environmentally caused reduced penetrance [1].

We began to wonder, however, what the power was of this segregation analysis (in the sense of type II statistical error) or, for that matter, any segregation analysis. Specifically, we wanted to ask the following questions: (1) Can this segregation analysis method distinguish between a fully-penetrant two-locus model and single-locus Mendelian models with environmentally caused reduced penetrance? (2) If not, at what point do the fully-penetrant two-locus models and the single-locus models with reduced penetrance become indistinguishable? (3) If the reduced-penetrance, single-locus models are indistinguishable from the fully penetrant two-locus models, will taking account of environmentally caused reduced penetrance in the two-locus models improve the ability of the method to distinguish models? (4) How much of a role does the ascertainment probability play in the segregation analysis?

The segregation analysis method that we used seemed well suited to answer these questions for two important reasons: it maximizes the likelihood of the data with respect to only two parameters, and it assumes a mode of inheritance more complex than single-locus Mendelian but one which needs less parameterization than the more complex segregation analysis programs. Also, the biological importance of two-locus models has become widely recognized recently [5]. There is good evidence that coeliac disease is the result of two recessive loci [1, 4, 6, 7]. In addition, hyperlipoproteinemia is the result of two loci [8], and Graves disease [9] and insulin-dependent diabetes mellitus [10–12] have both been suggested as candidates for a two-locus mode of inheritance. It has also been suggested that other HLA-related diseases, especially autoimmune diseases, may be the result of a locus within the HLA system and a second, non-HLA-linked locus [13].

METHODS

Throughout what follows, we will distinguish between the *generating* model (used to generate the data) and the *assumed* model (used to analyze the data). The simulation and analysis take enormous amounts of computer time. For that reason, we examined only the double-recessive (RR) model as the assumed model in this report. In this model, one must have a double dose of the disease alleles at both loci in order to be affected.

Simulation

Family data were simulated according to the following scheme:

(1) Choose at random a mating type capable of producing affected children, weighted by the frequency of the mating type in the population (which is a function of the gene frequencies of the disease alleles).

(2) Choose a family size according to a distribution. The distribution of family sizes in the population (a fitted negative binomial) was that described in [14]. While Ewens [15] and Morton [16] showed that family-size distribution should not affect the segregation analysis, the distribution of family sizes can affect power calculations because a data set containing 100 two-child families contains less information than, say, 100 five-child families. Therefore, we chose to use a realistic distribution of familv sizes in the simulation.

(3) For each child, determine whether that child is genetically affected, according to the segregation ratio for the chosen mating type. If the child is (genetically) affected, determine whether the child is phenotypically affected (i.e., allowing for reduced penetrance, if appropriate). If there are no affected children, go back to step (1).

(4) For each phenotypically affected child, determine whether that child is a proband, according to the predetermined probability of ascertainment, which was an input parameter. If there are no probands, the family is discarded.

For most analyses, 100 families were simulated per data set. Although 50-family data sets produced very similar results, there were more frequent terminations of the minimization routine due to round-off errors and somewhat larger standard deviations for the parameters. One hundred data sets of 100 families each were simulated in each computer run.

Segregation Analysis

The segregation analysis method used has been described [1]. It is modified from a method originally proposed by Elandt-Johnson [2] and calculates the likelihood of a nuclear family [$L(t, \pi)$]. It has the following form:

$$L(t, \pi) = \frac{\Sigma M_i (\alpha_i t)^a (1 - \alpha_i t)^b \pi^c (1 - \pi)^{(a-c)}}{1 - \Sigma M_i (1 - \pi \alpha_i t)^{(a+b)}} ,$$

where i indicates the ith mating type, M_i = probability of mating type i, α_i = segregation ratio for mating type i, t = the test parameter such that for the null hypothesis, $t = 1$, π = ascertainment probability, a = number of affected offspring, b = number of unaffected offspring, and c = number of probands. The log of $L(t, \pi)$ is summed over the individual families.

For the unrestricted hypothesis, the log likelihood was maximized with respect to two parameters: the test parameter (t) and the ascertainment probability (π). The likelihood ratio test was used to determine whether a data set supported or rejected the assumed model. For the null hypothesis, the log likelihood was maximized with respect to π alone and with t set to 1. The log-likelihood ratio multiplied by two (LR) was then computed: $LR = 2\{\mathrm{Log_e}[\mathrm{MAX}\ L(t, \pi)] - \mathrm{Log_e}[\mathrm{MAX}\ L(1, \pi)]\}$, which is distributed as a chi-square. Any data set where the LR exceeded 3.84 was treated as not supporting the assumed model (corresponding to a chi-square significance level of .05). The means and standard deviations of $\hat{t}$ and or $\hat{\pi}$ were computed for the 100 data sets.

Input parameters for the simulation (i.e., the generating model) included the allele frequencies, π, mode of inheritance, and penetrance. For the assumed model, only the mode of inheritance and the assumed penetrance were specified.

No attempt was made to estimate the gene frequencies of the disease alleles at the two disease loci simultaneously with $\hat{\pi}$ and $\hat{t}$. Instead, all combinations of gene frequencies that led to population prevalences in the range of five-times-greater to five-times-less than the "observed" (i.e., the prevalence specified in the simulation) were examined.

All analyses assumed the double-recessive (RR) model. Among the models used to simulate the data were: RR, dominant-recessive (DR), single-locus recessive, and single-locus dominant, both with and without reduced penetrance. "Reduced penetrance" in this case is assumed to result solely from environmental causes.

When data were generated under the single-locus dominant and recessive models, the generating gene frequency was chosen to give a population prevalence of about 1:1,600. The initial calculations showed that the generating value of π made little difference in the final outcome of the segregation analysis. Therefore, the generating model π, except where noted, was set at .5.

The likelihood was maximized using the IMSL program ZXMIN. The maximization was constrained so that $\hat{\pi}$ could not go below .001 nor above .999. Similarly, $\hat{t}$ was constrained to be between .1 and 2.0. [While there is no reason why $\hat{t}$ could not go above 2.0, a $\hat{t}$ greater than about 1.6 invariably led to rejection. The limit was set to reduce computer time. In addition, with $\alpha_i > .5$, a t greater than 2.0 leads to meaningless likelihoods due to the $(1 - \alpha_i\, t)$ term in the likelihood equation. This occurs only in the event of rare mating types, but these meaningless likelihoods must be trapped.]

Ascertainment Correction by Conditioning on Probands

Cannings and Thompson [17] suggested an ascertainment correction that does not require the simultaneous estimation of π. This method conditions the likelihood of the family on the probands. Specifically, the likelihood of a nuclear family takes the form:

$$L(t) = \frac{\Sigma\, M_i\, (\alpha_i\, t)^a\, (1 - \alpha_i\, t)^b}{\Sigma\, M_i\, (\alpha_i\, t)^c} \ .$$

(See above for definitions of the variables.)

We did several calculations varying the value of π used to generate the data in order to see how well this ascertainment correction worked.

RESULTS

First, a summary of the results: (1) The segregation analysis has sufficient power to distinguish between the fully-penetrant RR model and the fully-penetrant single-locus dominant and recessive models. (2) The method can also distinguish fairly well between the DR and RR models, even when one does not take into account the population prevalence. (3) The method has much less power to distinguish between the fully-penetrant RR model and the single-locus models with reduced penetrance. (4) The estimates of ascertainment probability, π, were robust, regardless of the model under which the data were generated. (5) The Cannings-Thompson approach to ascertainment correction worked well only when the π used to generate the data was less than .1.

Table 1 summarizes results when data are both generated and analyzed under an RR model assuming different values of the generating π. The actual rejection rate, or type I error, corresponds closely to the nominal rejection rate of .05. The mean $\hat{t}$ for all the runs is very close to 1.0 and the mean of the estimated π is close to that used to generate the data.

Table 2 summarizes the results when data were generated under a single-locus dominant, single-locus recessive, or DR model. When the penetrance of the generating model was 1, the analysis rejected almost all combinations of gene

TABLE 1

RESULTS OF ANALYSES OF DATA GENERATED UNDER THE RR MODEL USING DIFFERENT VALUES OF THE ASCERTAINMENT PROBABILITY AND ANALYZED UNDER THE RR MODEL

Generating π	% rejected (type-I error)	Mean $\hat{t}$ ± SD	Mean $\hat{\pi}$ ± SD	Range of mean $\hat{\pi}$*
.1	5	0.99 ± .23	.09 ± .07	.09–.10
.5	5	1.01 ± .23	.49 ± .11	.47–.51
.9	6	0.99 ± .26	.89 ± .06	.89–.90

* As a function of gene frequency.

frequencies. For data generated under the single-locus models, the mean $\hat{t}$ was about 1.9 ($\hat{t}$ was constrained to be below 2.0 in the maximization). Note, however, that the estimates of π, the ascertainment probability, are quite robust—about .5 for the recessive model, .43 for the dominant, and .40 for the DR model.

Table 2 also shows the effect of generating data under the simple Mendelian models with reduced penetrance. Penetrances between .5 and .1 were examined. Discrimination between these models and the fully penetrant RR model were considerably worse than when data were generated with fully-penetrant single-locus models. With reduced penetrance, the power to reject the RR model varied between about 17% and 96%. However, estimates of π remained surprisingly robust, varying between .43 and .57, when the generating π was .5. (While we have not tested whether the estimates of π would be worse if more extreme generating values of π were chosen, results in table 1 indicate that the estimates of π would be equally robust.)

Table 3 shows the effect of reducing the penetrance in the *assumed* model. Power improved substantially when the penetrance of the assumed model matched that of the generating model. The power appears to be a function of the ratio of

TABLE 2

RESULTS OF ANALYSES WHERE DATA WERE GENERATED UNDER A SINGLE-LOCUS DOMINANT OR RECESSIVE OR A DR MODEL AND ANALYZED ASSUMING AN RR MODEL WITH FULL PENETRANCE

GENERATING MODEL		ANALYSIS		
Model	Penetrance	Mean $\hat{t}$ ± SD	Mean $\hat{\pi}$ ± SD	% REJECTED
Dominant	1.0	1.92 ± .10	.43 ± .05	100
	.5	1.37 ± .19	.43 ± .08	59
	.25	0.91 ± .22	.43 ± .11	17
	.1	0.43 ± .16	.47 ± .21	81
Recessive	1.00	1.95 ± .11	.49 ± .11	100
	.5	1.24 ± .29	.48 ± .12	23
	.25	0.63 ± .22	.47 ± .18	37
	.1	0.27 ± .13	.46 ± .26	96
DR	1.0	0.61 ± .10	.40 ± .09	94

NOTE: Generating π was .5.

TABLE 3

RESULTS OF ANALYSES WHERE DATA WERE GENERATED UNDER A SINGLE-LOCUS DOMINANT OR RECESSIVE WITH REDUCED PENETRANCE AND ANALYZED ASSUMING AN RR MODEL WITH REDUCED PENETRANCE

Generating model	Generating penetrance	Assumed penetrance	% rejected	Mean $\hat{t}$ ± SD	Mean $\hat{\pi}$ ± SD
Recessive	.5	1.0	23	1.24 ± .29	.48 ± .12
	.5	.5	76	1.79 ± .30	.42 ± .11
	.1	.5	52	0.47 ± .18	.56 ± .21
Dominant	.5	1.0	59	1.37 ± .19	.43 ± .08
	.5	.5	100	1.96 ± .21	.45 ± .11
	.1	.5	14	0.71 ± .24	.51 ± .17

NOTE: Generating π was .5.

the generating and the assumed penetrances, that is, the power to reject the RR model is at a maximum when the ratio of the penetrances is unity.*

Table 4 shows the results of the calculations using the proband-conditioning ascertainment correction. This approach worked quite well as long as the value of π used to generate the model was less than about .1, a condition approaching single ascertainment [18].

The Cannings and Thompson approach, then, is biased. This bias has the effect of lowering the estimated value of $\hat{t}$, the test parameter, as the generating value of π increases. In the extreme, as π goes to 1, $\hat{t}$ goes to zero.

DISCUSSION

One can draw several conclusions from this analysis: (1) The Elandt-Johnson method can readily distinguish between the fully penetrant RR model and fully penetrant dominant or recessive inheritance or DR inheritance. (2) As the penetrance of the generating model goes down, the ability to distinguish between the single-locus Mendelian modes of inheritance and fully penetrant RR model also diminishes. It must be emphasized that the reduction in penetrance in this case is that caused solely by environmental influences. (3) Including reduced penetrance for the assumed model appears in large part to correct the effects of the reduced penetrance in the generating model. If there is a reasonable estimate to the environmentally caused reduced penetrance, then the method can distinguish between the one- and two-locus models. When the generating model was dominant and the penetrance was .5, if the analyzing penetrance was also .5, 100% of the data sets rejected the RR model at gene frequencies leading to compatible population prevalences. When a penetrance of 1 was used for the analyzing model, only 59% of the data sets rejected the RR model. Similarly, when the recessive model with 50% penetrance was used to generate the data, assuming a penetrance of 50% for the analyzing model led to 76% rejection, as opposed to only 23% rejection when a

* We were unable, for these studies, to use an assumed penetrance of less than .5. At penetrances lower than that, the maximization routine started to fail because of excessive round-off errors, even though all calculations were done in double precision (56 bits of accuracy).

fully penetrant RR model was assumed. (4) The estimates of the ascertainment probability are in reasonable agreement with the "true" values. The estimates of π are quite robust and appear to be almost independent of the models used to generate or analyze the data, at least when the generating π equals .5. Since calculations showed $\hat{\pi}$ to be consistently close to the generating π, the method is probably robust over the full range of π. (5) The proband-conditioning ascertainment correction works well if the ascertainment probability is less than about .1. Above a π of .1, substantial bias occurs in $\hat{t}$, the test parameter. This ascertainment correction does have the advantage that there is one less parameter in the maximization.

The power estimates may actually be better than those shown in table 2. For example, suppose data are generated under a single-locus model with reduced penetrance. For a given data set, most of the chi-square values for the different gene frequency combinations might not reject the model. However, if $\hat{t}$ is consistently different from 1, and the trend of $\hat{t}$ as a function of the different gene frequencies is in the direction of, say, a high gene frequency for one of the loci and a low frequency at the other locus, this might argue against the two-locus hypothesis, even though significance for any one gene frequency combination is not reached.

There are undoubtedly specific generating penetrances with the single-locus models that would give results that are indistinguishable from the fully-penetrant RR model. Those penetrance values would appear to be between .25 and .5. Figures 1 and 2 show graphs of the mean $\hat{t}$ as a function of the generating penetrance. The graphs indicate that, for the gene frequencies examined, a "true" penetrance of about .3 with dominant inheritance would lead to a complete inability of the method to distinguish between the RR model and the single-locus dominant model. The corresponding critical penetrance value for the recessive model is about .4.

It would appear from table 3 that taking into account the environmentally caused penetrance restores the power to distinguish models that was lost when data were generated under a single-locus model with reduced penetrance. Such environmental penetrance would be reflected in the monozygotic twin concordance rate. Unfortunately, reliable monozygotic twin rates are often difficult to obtain.

Figure 3 shows a plot of the mean value of $\hat{t}$ as a function of the π used to generate the data, when analysis was done using the proband-conditioning as-

TABLE 4

RESULTS OF ANALYSES OF DATA GENERATED UNDER THE RR MODEL USING DIFFERENT VALUES OF THE ASCERTAINMENT PROBABILITY WITH THE PROBAND-CONDITIONING ASCERTAINMENT CORRECTION AND ANALYZED UNDER THE RR MODEL

Generating π	% rejected	Mean $\hat{t}$ ± SD
.05	3	.95 ± .18
.1	6	.91 ± .19
.3	38	.72 ± .18
.5	78	.52 ± .16

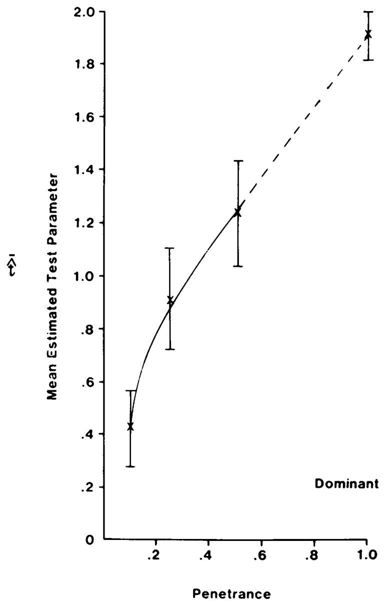

FIG. 1.—The mean estimated test parameter, $\hat{t}$, vs. the generating model penetrance, where the generating model was the single-locus dominant. *Dashed line* indicates that, since the value of $\hat{t}$ was not permitted to go above 2.0, the latter part of the curve is inaccurate. *Error bars* indicate plus and minus 1 SD.

certainment correction. The mean of $\hat{t}$ appears to be a linear function of the π used to generate the data. It should be noted that Cannings and Thompson derived the ascertainment correction under the assumption of one proband per family. Therefore, assuming anything other than single ascertainment violates that assumption. We wanted to see to what extent the method was biased at different generating ascertainment probabilities. (Further work on the Cannings-Thompson approach will appear in a paper by M. Boenhke and D. A. Greenberg; in preparation.)

The segregation analysis done here is very simple compared to some of the methods that have been reported. It consists of only two parameters: t, the test parameter, and π, the ascertainment probability. Under circumstances where one can use the proband-conditioning correction for the ascertainment probability, the method reduces to only one parameter. Despite its relative simplicity and restrictiveness, the *power* of the method to distinguish between genetic models is limited. Were one to try to estimate other parameters (such as penetrance or gene frequency), the power of the method to distinguish between competing models would be even less. This analysis points out the need to do simulation studies to establish the power of a segregation method. It also points out that a segregation analysis represents only one possible approach. Other ways of examining data [1, 5] are essential before one can reasonably expect to determine mode of inheritance.

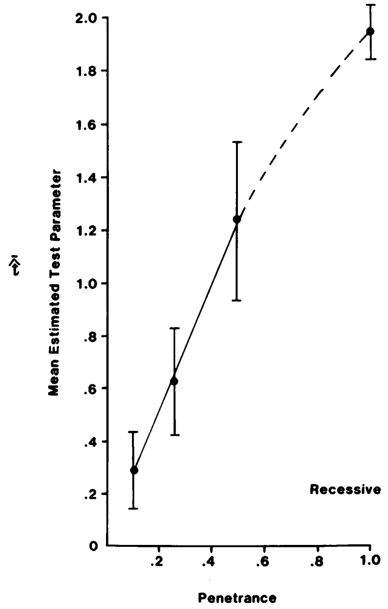

FIG. 2.—Same as in figure 1, but for the single-locus recessive model

One piece of data that was not taken advantage of in the current analysis was population prevalence. The limits on the population prevalence were deliberately set wide. All gene frequencies leading to population prevalences between five-times-greater and five-times-less than the "observed" were examined. In practice, population prevalences would usually be more restrictive than those limits. In the case of coeliac disease, where this analysis method was applied to data from

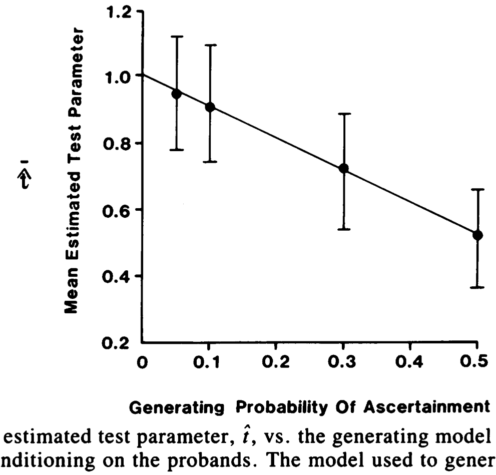

FIG. 3.—The mean estimated test parameter, $\hat{t}$, vs. the generating model π when the calculations were performed by conditioning on the probands. The model used to generate and analyze the data was the RR model.

41 families [1], reported population prevalence did vary by a factor of 5 in each direction (from about 1:6,000 to 1:300). However, even in that case, the population prevalence contributed to the argument against the dominant-recessive model.

The studies reported here were extremely time-consuming, both in human time and in computer time. Yet such studies are an important way of testing the tools that geneticists use.

ACKNOWLEDGMENTS

I wish to thank Drs. Kenneth Lange and Susan Hodge for helpful discussion and for criticizing the manuscript.

REFERENCES

1. GREENBERG DA, LANGE KL: A maximum likelihood test of the two-locus model for coeliac disease. *Am J Med Genet* 12:75–82, 1982
2. ELANDT-JOHNSON RC: Segregation analysis for complex modes of inheritance. *Am J Hum Genet* 22:129–194, 1970
3. PEÑA AS, MANN DL, HAGUE WE, ET AL.: Genetic basis of gluten sensitive enteropathy. *Gastroenterology* 75:230–235, 1978
4. KAGNOFF MF: Two genetic loci control the murine immune response to A-gliadin, a wheat protein that activates coeliac sprue. *Nature* 296:158, 1982
5. GREENBERG DA: A simple method for testing two-locus models of inheritance. *Am J Hum Genet* 33:519–530, 1981
6. GREENBERG DA, ROTTER JI: Two locus models for gluten sensitive enteropathy: population genetic considerations, *Am J Med Genet* 8:205–214, 1981
7. GREENBERG DA, HODGE SE, ROTTER JI: Evidence for recessive and against dominant inheritance at the HLA-"linked" locus in coeliac disease. *Am J Hum Genet* 34:263–277, 1982
8. UTERMANN G, LANGENBECK U, BEISIEGEL U, WEBER W: Genetics of apolipoprotein E-system in man. *Am J Hum Genet* 32:339–347, 1980
9. UNO H, SASAZUKI T, TAMAI H, MATSUMOTO H: Two major genes linked to HLA and Gm control susceptibility to Graves disease. *Nature* 292:268–270, 1981
10. THOMSON G: A two-locus model for juvenile diabetes. *Ann Hum Genet* 43:383–398, 1980
11. HODGE SE, ANDERSON CE, NEISWANGER K, ET AL.: Close genetic linkage between diabetes mellitus and Kidd blood group. *Lancet* ii:893–895, 1981
12. NAKAO Y, MATSUMOTO H, MIYAZAKI T, ET AL.: IgG heavy chain (Gm) allotypes and immune response to insulin-requiring diabetes mellitus. *N Engl J Med* 304:407–409, 1981
13. GREENBERG DA, ANDERSON CE: The search for heterogeneity in insulin dependent diabetes mellitus: evidence for familial and nonfamilial forms. *Am J Med Genet* 14:487–499, 1983
14. CAVALLI-SFORZA LL, BODMER WF: *The Genetics of Human Populations*. San Francisco, W. H. Freeman, 1971, pp 310–313
15. EWENS WJ: Aspects of parameter estimation in ascertainment sampling schemes. *Am J Hum Genet* 34:853–865, 1982
16. MORTON NE: Trials of segregation analysis by deterministic and macro simulation. *Am J Hum Genet* 34:187A, 1982
17. CANNINGS C, THOMPSON EA: Ascertainment in the sequential sampling of pedigrees. *Clin Genet* 12:208–212, 1977
18. MORTON NE: *Outline of Genetic Epidemiology*. Basel, Switzerland, S. Karger, 1982