Genetic Epidemiology 20:397–400 (2001)

LETTER TO THE EDITOR

On the Total Expected Study Cost in Two-Stage Genome-Wide Search Designs for Linkage Analysis Using the Mean Test for Affected Sib Pairs

To the Editor: Recently, Guo and Elston [2000] studied optimal two-stage global search designs for linkage analysis using affected sib pairs (ASP). Their method is an extension of the approach proposed by Elston et al. [1996] who have shown that typically half the cost can be saved compared to a one-stage design. The aim of this letter is to show that the total expected study cost is not adequately reflected in their approach. It is furthermore sketched how this problem could be dealt with. Finally, we discuss the term "power to detect linkage," which is essential for an appropriate calculation of the required sample size.

Based on the model by Guo and Elston [2000], the notation required to introduce a cost function for a two-stage design is as follows: The total study cost generally depends on the number of individuals to be genotyped and phenotyped. If n independent ASP are recruited without parents, this number is $2n$. The study cost also depends on the cost per marker (C_1) and the phenotyping cost per individual (C_2). Then, the total expected cost is the sum of the phenotyping cost $2n\,C_2$, the cost $2n\,m\,C_1$ of genotyping $2n$ individuals with m markers at the first stage, and the cost for genotyping all $2n$ individuals at $2k$ additional markers at the second stage for those first stage markers that demonstrate P values $<< \alpha^*$. $2k$ additional markers are typed at those positions that are either false or true positive. Guo and Elston [2000, eq. (2)] presented the following equation to denote the total expected study cost in units of C_1

$$C = 2n\left\{\frac{C_2}{C_1} + m + 2k\left[\alpha^* m + (1-\beta)d\right]\right\}, \tag{1}$$

Contract grant sponsor: Deutsche Forschungsgemeinschaft; Contract grant sponsor: Cedars-Sinai Board of Governor Chair in Medical Genetics.

where β is the type II error rate. It is easily seen that $\alpha^* m$ aims to measure the number of markers that are false positive at the first stage. However, m denotes the total number of first stage markers, thus including true positives. This means that $\alpha^* m$ should be reduced by the number of marker loci from the first stage that correctly show linkage. The reader should note that $(1 - \beta)\, d$ is the expression for the number of true disease loci that are typed in the second stage. The cost function (1), however, ignores that there may be further linked markers around the disease locus.

This can be shown by assuming a disease locus to be completely linked, i.e., recombination fraction $\theta = 0$, to a first stage marker, say M_d. From the construction principle of the procedure, $2k$ additional markers are typed around M_d with probability $\geq 1 - \beta$. The true effect δ_d at M_d is

$$\delta_d = \frac{2\lambda_S - \lambda_O - 1}{4\lambda_S}$$

since $\theta = 0$. Hence, the effect at those first stage markers M_l and M_r flanking M_d is $\delta_l = \delta_r = ([\theta^{*2} + (1 - \theta^*)^2] - 1)\delta_d$, with $\theta^* = K(2x^*)$ for a map function K. Thus, $2k$ markers are typed in stage 2 around both M_l and M_r for large n with probability

$$1 - \beta_l = 1 - \Phi\left(\frac{\frac{1}{2} z_{\alpha^*} - \delta_l \sqrt{2n}}{\sqrt{\frac{1}{4} - \delta_l^2}} \right), \tag{2}$$

where Φ is the standard normal distribution function. Note that in the model of Guo and Elston [2000] $2k$ stage 2 markers are typed around any marker other than the most closely linked marker at the first stage with probability $\alpha^* << 1 - \beta_l$. This increase in probability (2) affects both the total expected study cost and the overall power.

TOTAL EXPECTED STUDY COST

The total expected cost obtained with the approach of Guo and Elston [2000] is more appropriately reflected by

$$C = 2n\left\{ \frac{C_2}{C_1} + m + 2k\left[\alpha^* (m - \sum_{i=1}^{d} l_i) + \sum_{i=1}^{d} \sum_{j=1}^{l_i} (1 - \beta_{ij}) \right] \right\}, \tag{3}$$

where l_i is the number of first stage markers linked to disease locus i. For each first stage marker j linked to disease locus i, $1 - \beta_{ij}$ is the probability that $2k$ second stage markers around it are typed. $m - \Sigma_{i=1}^{d} l_i$ is the number of stage one markers that are not linked to any disease locus with effects λ_s and λ_o.

The increase in total expected cost was compared between cost function (3) and cost function (1) using the mean test and the parameter constellations of the optimal two-stage design [Guo and Elston, 2000, Table I]. It was assumed that there is one

TABLE I. Total Expected Cost for the Two-Stage Design Proposed by Guo and Elston [2000] and Increase of Total Expected Cost (in percent) According to Cost Function (3)*

		λ[a]			
C_2/C_1[b]	d[c]	1.2	1.5	2	5
0	1	492,656 (8.45)	121,894 (8.40)	53,210 (8.24)	19,696 (7.80)
	2	505,251 (16.48)	124,985 (16.39)	54,548 (16.08)	20,176 (15.23)
100	1	875,564 (11.08)	215,866 (11.05)	93,699 (10.98)	34,090 (10.46)
	2	892,460 (21.74)	220,003 (21.69)	95,475 (21.55)	34,723 (20.55)
200	1	1,225,842 (8.05)	301,762 (8.00)	130,635 (7.95)	47,151 (7.77)
	2	1,242,575 (15.88)	305,871 (15.78)	132,401 (15.68)	47,775 (15.33)

*Number of first- and second-stage markers were calculated for the optimal two-stage design [Guo and Elston, 2000] using $M = 33$, $\alpha = 0.0001$, $1 - \beta = 1 - \beta^* = 0.8$. It was assumed that the distance between the disease locus and the outermost first stage marker locus was 80 cM. Kosambi was used as map function.
[a]Linked disease locus risk ratio for siblings and for parents-offspring.
[b]Phenotyping-genotyping cost ratio.
[c]Number of disease loci.

disease locus, which is located in the middle of a chromosome of a length of about 100 and 160 cM, respectively, while scanning the whole genome for susceptibility loci. If the distance between the disease locus and the outermost first stage markers is 80 cM, the total expected cost is between 2.83 and 8.60% greater if cost function (3) is used instead of (1) (detailed results not shown). Thus, the difference is negligible for the models considered by Guo and Elston [2000].

We also studied the effect of relative risk ratio λ, the genotype-phenotype cost ratio C_2/C_1, and the number of disease loci d on the total expected cost. Table I displays both the total expected cost and the increase in total expected cost for each optimal two-stage design assuming an additive genetic model, i.e., $\lambda_s = \lambda_o$, a global significance level $\alpha = 0.0001$, which approximately corresponds to LOD = 3, and 80% power for first- and second-stage markers. Furthermore, we considered a distance of 80 cM between the disease locus and the outermost first stage marker on a chromosome and used the Kosambi map function, while scanning the whole genome for susceptibility loci. Markers were assumed to be completely informative. The increase in the total expected cost on use of (3) is relevant for the displayed models. As expected, the ratio of the total expected cost calculated by (3) compared with (1) increases with the number of disease loci d. Furthermore, it is generally higher for a greater phenotyping-genotyping cost ratio C_2/C_1. It also increases with decreasing significance level (results not shown). Nevertheless, the cost function (1) provided by Guo and Elston [2000] might serve as a first approximation to the total expected cost using an optimal two-stage design for ASP.

POWER

As shown before, the increase in probability (2) affects not only the total expected study cost but also the overall power. The term “power to detect linkage” can be used in at least two different ways.

1. A linkage to a disease locus is found, only if the marker next to the disease locus is significant.

2. A linkage to a disease locus is found, if any marker linked to the disease locus is significant.

In situation 1, the probability that the *P* value of a first stage marker linked to the disease locus remains under α^* is greater than α^*. This results in an inflation of the type I error at the first stage. Similar arguments can be applied to show that the overall type I error is greater than α. Here, the total expected study cost might have to be increased to maintain the overall significance level α.

In situation 2, the same arguments as above can be used to show that the power to detect linkage is greater than $1 - \beta$. Here, the total expected study cost could be reduced upon calculation of the true power. The computation of the true overall power and the true overall significance level depends, however, on the joint distribution of the first and second stage markers. This would require the calculation of high dimensional standard normal distribution functions.

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (F.G.) and by the Cedars-Sinai Board of Governor Chair in Medical Genetics (X.G.).

Andreas Ziegler
Inke Böddeker
Frank Geller
Hans-Helge Müller
Institute of Medical Biometry and Epidemiology
Philipps-University of Marburg
Marburg, Germany

Xiuqing Guo
Division of Medical Genetics
Department of Medicine and Pediatrics
Spielberg Pediatrics Research Center
Burns and Allen Research Institute
Cedars-Sinai Medical Center
Los Angeles, California

REFERENCES

Elston RC, Guo X, Williams LV. 1996. Two-stage global search designs for linkage analysis using pairs of affected relatives. Genet Epidemiol 13:535–58.

Guo X, Elston RC. 2000. Two-stage global search designs for linkage analysis I: use of the mean statistic for affected sib pairs. Genet Epidemiol 18:97–110.