Published in final edited form as:

Genet Epidemiol. 1993 ; 10(6): 465–470. doi:10.1002/gepi.1370100623.

A Bootstrap Approach to Estimating Power for Linkage Heterogeneity

Suzanne M. Leal and Jurg Ott

Department of Psychiatry, Columbia University, New York

Abstract

We examined the power of detecting linkage heterogeneity when the null hypothesis is that all families are linked to one locus (A) and the two alternative hypotheses are either 1) a proportion of the families are linked to locus A and the remaining families are linked to a second locus B or 2) a proportion of the families are linked to locus A or B and a third proportion of the families are unlinked to either locus. The power of detecting linkage heterogeneity is estimated for various proportions of families linked to loci A, B or unlinked to either locus (sampling under the alternative hypothesis). To estimate the significance level, the data set is sampled under the null hypothesis. For sampling under both hypotheses, a bootstrap approach is employed, sampling the simulated pedigrees with replacement.

The power to detect linkage heterogeneity is strongest when the recombination fraction is 0 and equal proportions of the families are linked to loci A and B. The power decreases as the recombination fraction increases, the proportion of unlinked families increases and the disparity between the proportion of the families linked to either locus A or B increases. In the data set of 32 Duke Familial Alzheimer Disease families, when equal proportions of families are linked to loci A and B, the power to detect linkage heterogeneity is 0.94 using a likelihood ratio criterion of 10: 1. The p value that corresponds to the likelihood ratio of 10:1 is estimated as 0.013 with a 95% confidence interval for p ranging from 0.012 to 0.014.

Keywords

linkage heterogeneity; power analysis; bootstrap sampling

INTRODUCTION

Genetic heterogeneity involving two or more loci has been demonstrated in a number of genetic diseases including Charcot-Marie-Tooth disease [Bird et al., 1983; Vance et al., 1989] and tuberous sclerosis [Fryer et al., 1987; Kandt et al., 1992]. To assess the power of detecting locus heterogeneity and the significance level for a given likelihood ratio, we developed a computer program that uses a resampling strategy described by Tenvilliger and Ott [1992], which is modelled after the multisample bootstrap of Efron [1982].

METHODS

The following is a list of definitions to clarify the text:

θ = recombination fraction in generated data

n = number of families available for study

r = number of simulated replicates for each family

b = number of bootstrap samples

t = number of recombination fraction values at which the lod score was evaluated

S_i = $r \times t$ matrix of lod scores for the i^{th} family when the marker and disease loci are generated linked ($\theta < 0.5$)

U_i = $r \times t$ matrix of lod scores for the i^{th} family when the marker and disease loci are generated unlinked ($\theta = 0.5$)

The ability to detect genetic heterogeneity was investigated when the null hypothesis (H_0) is that all families are linked to one locus (A) and the two alternative hypotheses are either 1) a proportion α_1 of the families are linked to locus A and the remaining proportion $1-\alpha_1$ of the families are linked to locus B (Hypothesis H_1) or 2) a proportion α_1 of the families are linked to locus A, a second proportion α_2 of families are linked to locus B, and a third proportion $1-\alpha_1-\alpha_2$ of the families are unlinked to either locus (Hypothesis H_2). Loci A and B must be unlinked, but need not be located on two different chromosomes. The assumptions are made that the two disease loci A and B have identical parameters and that the recombination fractions between the marker and disease loci are the same. To test the null hypothesis versus each of the alternative hypotheses, we use likelihood ratio tests. Since the null distribution of the test statistic is not known, its significance level is estimated empirically.

Random sampling.

Two sets of marker genotype data for each of the n families are generated conditional on the observed disease phenotype. The first set of n families is generated r times with the marker locus linked ($\theta < 0.5$) to the disease locus (either locus A or B, since they have the same parameters) and in the second set of data with a marker locus unlinked to any disease locus ($\theta = 0.5$). The lod scores are evaluated using the true parameter values at t recombination fraction values for each of the generated pedigrees. That way, for families $1, \dots, n$, two sets of $r \times t$ matrices are created $S_1 \dots S_n$ and $U_1 \dots U_n$, where the rows (r) refer to the i^{th} family's replicates $1, \dots, r$, and the columns (t) correspond to the lod scores for each replicate evaluated at specific values of the recombination fraction. Matrix S_i contains the lod scores for the i^{th} family when the marker and disease loci were generated linked ($\theta < 0.5$) and matrix U_i contains those when the marker genotypes are generated unlinked to the disease locus ($\theta = 0.5$).

Bootstrap resampling.

We now resample from the r randomly generated replicates. For family 1, we pick a random row from matrix $\mathbf{S}_1$ (lod scores under linkage) and a random row from matrix $\mathbf{U}_1$ (lod scores under no linkage). For family 2, we analogously pick rows from matrices $\mathbf{S}_2$ and $\mathbf{U}_2$, and so on for each of the n families.

Depending on whether resampling is done under the null or one of the alternative hypotheses, the resampled lod score array is interpreted differently. When resampling under H_0 the lod scores from the matrices $\mathbf{S}_1 \dots \mathbf{S}_n$ represent the lod score arrays in bootstrap sample 1 between disease and locus A; the lod scores from the matrices $\mathbf{U}_1 \dots \mathbf{U}_n$ represent the lod score arrays between disease and locus B. Under H_1 , each family is “linked” to locus A with probability p_1 and to locus B with probability p_2 . We randomly determine for the i^{th} family whether its disease locus is to be linked to A or B. When linkage is to be to locus A, the lod scores from matrix $\mathbf{S}_i$ are taken to represent the lod scores (for family i) between disease and locus A, and the lod scores from $\mathbf{U}_i$ are the lod scores between disease and locus B. When linkage is to be to locus B, lod scores from $\mathbf{S}_i$ represent the lod scores between disease and locus B, and lod scores from $\mathbf{U}_i$ are the lod scores between disease and locus A. Finally, under H_2 the situation is analogous to the one described for H_1 except that, when family i is determined to be unlinked to both loci A and B, two independent arrays from matrix $\mathbf{U}_i$ serve as the lod scores between disease and locus A and disease and locus B. This completes taking the first bootstrap sample. The whole resampling process is repeated for bootstrap sample number two and so on.

Once a bootstrap sample is completed, it is tested for linkage heterogeneity using an extended version of the admixture test [Ott, 1991]. For the admixture test the maximum likelihood under the null hypothesis is $\prod_{i=1}^n [L_i(\hat{\theta}_1)]$ and the maximum likelihoods under the

two alternative hypotheses are $\prod_{i=1}^n [\alpha_1 L_i(\hat{\theta}_1) + (1 - \alpha_1) L_i(\hat{\theta}_2)]$ (H_1), and

$\prod_{i=1}^n [\alpha_1 L_i(\hat{\theta}_1) + \alpha_2 L_i(\hat{\theta}_2) + (1 - \alpha_1 - \alpha_2) L_i(\frac{1}{2})]$ (H_2). As test criteria for determining

significance, we use likelihood ratios of 10:1 and 100:1. The bootstrap resampling and linkage heterogeneity testing are repeated b times. Significance level and power are estimated as the proportion of significant bootstrap samples, from sampling done under H_0 or one of H_1 or H_2 , respectively.

As an empirical example, the disease phenotypes in the 32 Duke familial Alzheimer families were taken as given and marker genotypes were generated as described above. A disease gene frequency of 0.01 was assumed. The penetrance is age dependent with no phenocopies, the age of onset distribution being assumed normally distributed (mean $\pm$ s.d. = 66.1 ± 10.3 yrs). These parameter estimates were used in the linkage analysis that suggested evidence of linkage of familial Alzheimer’s disease to chromosome 19 in this data set of families [Pericak-Vance et al., 1991]. A marker with four alleles of equal frequency was simulated for those individuals who were typed for at least one marker in the original data set (64% of

the family members); all other individuals were considered unknown at the marker locus. The two disease loci were modelled to have identical parameters. A total of $r = 500$ replicates were simulated at $\theta = 0.02$, $\theta = 0.10$, and $\theta = 0.50$. To simulate the marker genotypes, SLINK [Weeks et al., 1990] was used when the recombination fraction was less than 0.5 and, for increased speed, SIMULATE [Ott and Terwilliger, 1992] was used when the recombination fraction was equal to 0.5. MLINK [Lathrop et al., 1984] was used to evaluate the lod scores for each of the simulated pedigrees. To estimate the power of detecting linkage heterogeneity, $b = 500$ bootstrap samples were generated under the alternative hypothesis; and to accurately estimate a presumably small value of the significance level under H_0 , $b = 20,000$ bootstrap samples were generated under the null hypothesis. HOMOG3R [Ott, 1991] was used to test each bootstrap sample for linkage heterogeneity.

RESULTS AND DISCUSSION

Table I displays the results of the power analysis. The first three columns show the assumed proportions of families linked to locus A, locus B, and unlinked to both loci. The results are given for recombination fractions between the disease and marker loci of 0.02 and 0.10. For each recombination fraction, the power of detecting linkage heterogeneity is displayed for two criteria, likelihood ratios of 10: 1 and 100: 1. The first row (proportions 1, 0, and 0) presents the significance levels associated with the given likelihood ratio criteria.

As shown in Table I, a likelihood ratio criterion of 10:1 is not very stringent; it corresponds to a significance level somewhat larger than 0.01. This seems sufficient to test for heterogeneity if linkage to both loci has previously been established [Cavalli-Sforza and King, 1986]. However, if heterogeneity also implies presence of linkage to a locus not previously detected, the usual stringency for detecting linkage should be applied; that is, a likelihood ratio criterion of at least 100: 1.

For a given likelihood ratio criterion, the estimated significance level increases with increasing recombination fraction. Both increases (from 0.013 to 0.029 and from 0.0002 to 0.002) are highly significant ($p < 0.001$). Of course, the significance levels decrease as the likelihood ratio criterion increases.

The power to detect heterogeneity is largest when the proportion of families linked to A and B are equal, and the power decreases as the proportion of unlinked families increases. The maximum power (for proportions 0.5, 0.5, and 0) is estimated as 94% for $\theta = 0.02$ and a likelihood criterion of 10:1, but drops to 21% for $\theta = 0.10$ and a more stringent likelihood ratio criterion of 100: 1. These power figures are estimated quite accurately. For example, based on the estimate of $105/500 = 0.21$, the 95% confidence-interval for the power is (0.18, 0.25) and the confidence interval associated with $470/500 = 0.94$ is (0.92, 0.96).

We chose a bootstrap approach over using simulation alone, since each bootstrap sample can be created from the simulated pedigrees in a fraction of the time it would take to simulate and evaluate each replicate. The amount of computer time needed for analysis is greatly reduced using bootstrap resampling, thus making it more practical to undertake a given

analysis. Since for each bootstrap sample each family is sampled independently, and not as a set of families, additional family sets are made available without additional simulation. Given that all of the r replicates are different from each other, there are a total of r^n different bootstrap samples. In our case, when sampling under one of the hypotheses, we have $r^n = 500^{32} = 2.3 \times 10^{86}$ possible quasi-independent bootstrap samples. It would then be unlikely that any two of them are the same. Of course, particularly for small family types, some of the r replicates randomly generated may not be unique.

A large number of bootstrap samples can be chosen because bootstrap sampling is relatively fast. The number of replicates that should be generated for each family will depend on the size, structure, and number of families in the analysis. In theory the number of generated replicates should be large enough that the complete sample space has been covered; in practice, however, this is usually not feasible.

The computer program developed to carry out this analysis is available from the authors.

ACKNOWLEDGMENTS

We wish to thank Marcella Devoto, Marcy Speer, and Joseph Tenwilliger for their comments and suggestions. This work was supported by NIH grant HG00008.

REFERENCES

- Bird TD, Ott J, Giblett ER, Chance PF, Sumi SM, Krafi GH (1983): Genetic evidence for heterogeneity in Charcot-Marie-Tooth neuropathy (HMSN-Type 1). *Ann Neurol* 14:679–684. [PubMed: 6651251]
- Cavalli-Sforza LL, King M- C (1986): Detecting linkage for genetically heterogeneous diseases and detecting heterogeneity with linkage data. *Am J Hum Genet* 38:599–616.
- Efron B (1982): “The Jackknife, the Bootstrap and other Resampling Plans.” Philadelphia: SIAM.
- Fryer AE, Connor JM, Povey S, Yates JRW, Chalmer A, Fraser I, Yates AD, Osborne JP (1987): Evidence that the gene for tuberous sclerosis is on chromosome 9. *Lancet* 1:659–661. [PubMed: 2882085]
- Kandt RS, Haines JL, Smith M, Northrup H, Gardner RJM, Short MP, Dumars K, Roach ES, Steingold S, Wall S, Blanton SH, Flodman P, Kwiakowski DJ, Jewell A, Weber JL, Roses AD, Pericak-Vance MA (1992): Linkage of an important gene locus for tuberous sclerosis to a chromosome 16 marker for polycystic kidney disease. *Nat Genet* 2:37–41. [PubMed: 1303246]
- Lathrop GM, Lalouel JM, Julier C, Ott J (1984): Strategies for multilocus linkage analysis in humans. *Proc Natl Acad Sci USA* 81:3443–46. [PubMed: 6587361]
- Ott J (1991): “Analysis of Human Genetic Linkage.” Baltimore: Johns Hopkins University Press.
- Ott J, Tenwilliger J (1992): Assessing the evidence for linkage in psychiatric genetics. In Mendlewicz J, Hippus H (eds): “Genetic Research in Psychiatry.” New York: Springer, pp 245–249.
- Pericak-Vance MA, Bebout JL, Gaskell PC, Jr, Yamaoka LH, Hung W-Y, Alberts MJ, Walker AP, Bartlett RJ, Haynes CA, Welsh KA, Earl NL, Heyman A, Clark CM, Roses AD (1991): Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *Am J Hum Genet* 48: 1034–1050. [PubMed: 2035524]
- Tenwilliger JD, Ott J (1992): A multisample bootstrap approach to the estimation of maximized-over-models lod score distributions. *Cytogenet Cell Genet* 59: 142–144. [PubMed: 1737483]
- Vance JM, Nicholson GA, Yamaoka LH, Stajich J, Stewart CS, Speer MC, Hung WY, Roses AD, Barker D, Pericak-Vance MA (1989): Linkage of Charcot-Marie-Tooth neuropathy type I a to chromosome 17. *Exp Neurol* 104:186–189. [PubMed: 2707366]

Weeks DE, Ott J, Lathrop GM (1990): SLINK: A general simulation program for linkage analysis. Am J Hum Genet 47(suppl): A204.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table I.

Significance Level and Power to Detect Linkage Heterogeneity

Proportion of families			$\theta = 0.02$		$\theta = 0.10$	
Linked to A	Linked to B	Unlinked	LR ^a of 10:1	LR ^a of 100:1	LR ^a of 10:1	LR ^a of 100:1
<i>Significance level</i>						
1.0	0.0	0.0	0.013	0.0002	0.029	0.002
<i>Power</i>						
0.9	0.1	0.0	0.244	0.106	0.130	0.032
0.8	0.2	0.0	0.558	0.326	0.262	0.082
0.7	0.3	0.0	0.784	0.514	0.412	0.114
0.5	0.5	0.0	0.940	0.752	0.522	0.210
0.45	0.45	0.1	0.844	0.570	0.480	0.166
0.4	0.4	0.2	0.794	0.472	0.356	0.100

^a Likelihood ratio.