European Journal of Human Genetics (2006) 14, 1018–1026
© 2006 Nature Publishing Group All rights reserved 1018-4813/06 $30.00
www.nature.com/ejhg

ARTICLE

# Covariate-based linkage analysis: application of a propensity score as the single covariate consistently improves power to detect linkage

Betty Q Doan*[,1,2,5], Alexa JM Sorant[3], Constantine E Frangakis[4], Joan E Bailey-Wilson[2] and Yin Y Shugart[1]

*[1]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; [2]Statistical Genetics Section, Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD, USA; [3]Genometrics Section, Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD, USA; [4]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA*

**Successful identification of genetic risk loci for complex diseases has relied on the ability to minimize disease and genetic heterogeneity to increase the power to detect linkage. One means to account for disease heterogeneity is by incorporating covariate data. However, the inclusion of each covariate will add one degree of freedom to the allele sharing based linkage test, which may in fact decrease power. We explore the application of a propensity score, which is typically used in causal inference to combine multiple covariates into a single variable, as a means of allowing for multiple covariates with the addition of only one degree of freedom. In this study, binary trait data, simulated under various models involving genetic and environmental effects, were analyzed using a nonparametric linkage statistic implemented in LODPAL. Power and type I error rates were evaluated. Results suggest that the use of the propensity score to combine multiple covariates as a single covariate consistently improves the power compared to an analysis including no covariates, each covariate individually, or all covariates simultaneously. Type I error rates were inflated for analyses with covariates and increased with increasing number of covariates, but reduced to nominal rates with sample sizes of 1000 families. Therefore, we recommend using the propensity score as a single covariate in the linkage analysis of a trait suspected to be influenced by multiple covariates because of its potential to increase the power to detect linkage, while controlling for the increase in the type I error.**

*European Journal of Human Genetics* (2006) **14,** 1018–1026. doi:10.1038/sj.ejhg.5201650; published online 31 May 2006

**Keywords:** covariate; affected relative pairs; linkage analysis; simulation; propensity score; LODPAL

## Introduction

A variety of nonparametric methods have been developed that test for linkage of a disease gene to a marker locus (or loci) by detecting significant increases in identical by descent (IBD) allele sharing probabilities in pairs of affected relatives. Incorporating covariate information can improve their power, and several methods have been proposed to allow for disease heterogeneity by using covariate-dependent penetrances that define liability

*Correspondence: Dr BQ Doan, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Broadway Research Building, 733 North Broadway, Room 572, Baltimore, MD 21205, USA.
Tel: +1 410 502 7537; Fax: +1 410 502 7544;
E-mail: bdoan@jhmi.edu
[5]Current Address: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA

Received 20 October 2004; revised 13 March 2006; accepted 13 April 2006; published online 31 May 2006

classes, or covariate-dependent allele sharing proportions that define linked and unlinked subgroups.[1–7]

A unified approach for linkage detection that incorporates covariate data directly in a general conditional logistic model has also been proposed.[8,9] The likelihood ratio test of linkage is based on the relative risk of disease ($\lambda_{ir}$) and the IBD allele sharing probabilities ($f_{ir}$) for a given relative pair type ($r$) summed over all possible numbers of alleles ($i = 0$, 1, or 2) shared IBD. The likelihood ratio for a relative pair type, ($LR_r$), compares the observed allele sharing proportions ($\hat{f}_{ir}$) given the genotype marker data with the probabilities ($f_{ir}$) expected under the hypothesis of no linkage using

$$LR_r = \sum_{i=0,1,2} \lambda_{ir}\hat{f}_{ir} \Bigg/ \sum_{i=0,1,2} \lambda_{ir} f_{ir}.$$

The model is extended with a parameterization of the relative risk $\lambda_{ir}$ for a given relative pair sharing $i$ (0, 1, or 2) alleles IBD by a conditional logistic regression model allowing for $K$ covariates, where

$$\lambda_{ir} = \exp\,(\beta_i + \sum_{j=1}^{K} \delta_{ij} x_j),$$

$\beta_i$ is the logarithm of the disease relative risk of sharing $i$ alleles IBD with no covariate effects, and $\delta_{ij}$ is the contribution to the logarithm of the disease relative risk per unit increase in covariate $x_j$ while simultaneously controlling for all other covariates. The relative risks are constrained[10] so that only one additional parameter is estimated for each included covariate. Directly incorporating covariates into the likelihood ratio test allows for covariate-related locus heterogeneity and for the simultaneous estimation of both linkage and covariate effects. Its limitation, however, is that each additional covariate adds one degree of freedom to the linkage test. Consequently, use of covariates needs to balance the increase in power by accounting for heterogeneity with multiple covariates with its reduction due to estimating more parameters.

Several genome-wide linkage analyses using this approach have been successful in enhancing previously identified linkage signals when compared to linkage analyses without covariates.[8,9,11–15] Stepwise model building and principal component analyses can assess the relative importance of covariates and guide their selection, but they raise multiple testing concerns when many subsets are analyzed.[13] Unknown underlying covariate effects on the disease trait make an a *priori* selection of important covariates difficult. We propose to overcome these problems related to including multiple covariates by using a propensity score (PS), which collapses multiple covariate data into a single scalar variable.

Rosenbaum and Rubin[16] first described the PS in a causal inference analysis as a means to control for multiple covariate effects that could potentially bias assessments of treatment effect on outcomes when randomization experiments were not possible. In such a setting, the score is defined as the conditional probability of being assigned to a treatment (the causal event) given the covariate data, and in practice, can be estimated from observed covariate data with a logistic regression of treatment group assignment on the covariates. The score can then be used for matching, stratification, or regression adjustment to ensure that the covariates are balanced between the two treatment groups when outcomes between groups are compared. This is feasible because the PS has the balancing property that for groups of subjects with the same score, the treated and the untreated subjects have the same joint distribution of all the covariates that entered into its calculation.[16–18]

In the context of linkage analysis, the causal event of interest, namely, the genotype at a locus responsible for disease, is not directly observed because that locus is not known a *priori*. For this reason, the standard definition of a PS is not directly applicable, but it can be redefined to be the conditional probability of being affected given the covariates. Such a definition of a PS for predicting affection was also proposed by Rich,[19] but was not examined in practice. Here we propose to use the PS as a covariate in linkage analysis with an affected relative pair design.

After conditioning on the PS, the covariates no longer predict disease status because the balancing property ensures that the affected and unaffected individuals with the same score have the same joint distribution of all the covariates considered. Consequently, stratification with a single PS will capture similar information to that of multiple covariates without further increasing the degrees of freedom. Secondly, the PS can be interpreted as the predicted disease risk contributed by the covariates considered. Thus, disease mediated through pathways other than the covariates, such as through a genetic locus will be more easily detectable from subjects who both have low PS, and are more homogenous. With affected relative pairs, this means that a model that allows the excess IBD sharing to associate with low pair-specific PS strata (such as with pairs with the smallest sum) is expected to be more powerful than a model that does not.

We tested the application of the PS through a series of simulations of a genetic trait with underlying covariate effects. In our analysis, we first estimated the PS using both affected and unaffected individuals. Then we used the estimated PS as the single covariate in Olson's conditional logistic regression model[8] within the affected relative pair design, and compared its performance in linkage analyses incorporating no covariate, each covariate individually, and all covariates simultaneously. We seek answers to two questions: (1) does the inclusion of the PS as the single covariate improve the power to detect linkage over the inclusion of either covariate individually or of all the covariates simultaneously; and (2) how does its type I error rate compare?

## Materials and methods

### Simulation

Using G.A.S.P. v3.33,[20] a binary trait phenotype was simulated based on an effective disease penetrance (EDP) model for the underlying genetic trait locus and two covariate effects (one binary and one quantitative). Genotype data for linked ($\theta = 0.01$) and unlinked ($\theta = 0.5$) markers, each with eight equally frequent alleles, were also simulated. Two sets of genotype-specific baseline penetrance values defined dominant, codominant, and recessive genetic inheritance models based on a disease allele frequency ($q$) of 0.1% for a dominant or codominant disease allele or of 2% for a recessive disease allele.

The effective disease penetrance (EDP) was defined as the probability of affection ($A$) dependent on the baseline penetrance given genotype $G$ (11, 12, or 22) at the trait locus and on the binary ($B$) and quantitative ($Q$) covariates effects using the model:

$$EDP = P(A|G, B, Q) = \begin{cases} 0 & \text{if } 0 > \text{sum} \\ \text{sum} & \text{if } 0 < \text{sum} < 1 \\ 1 & \text{if sum} > 1 \end{cases},$$

with $\text{sum} = p_B(G) + C_B(G)^*B + C_Q(G)^*Q$ where: $p_B(G)$ is the baseline disease penetrance for genotype $G$; $B$ is an indicator of binary covariate exposure with probability $P_B$; $Q$ is a quantitative covariate (drawn from a standard normal distribution); and $C_B(G)$, $C_Q(G)$ are genotype dependent coefficients of the contribution to the effective disease penetrance for covariates $B$, $Q$, respectively.

A total of 200 nuclear families having four offspring (at least two affected) were ascertained for each sample, and 10 000 replicates were generated for each model. All trait, covariate, and marker data were available for all individuals. Covariate values were independent of genotypes and of each other.

### Propensity score calculation

The PS values were derived from a logistic regression of affection status using SAS v8.2. on the entire data set with covariate data, fitting the following logit model:

$$\text{logit}\left[Pr\ (\text{individual } i \text{ affected}\left|\underline{x_{ij}}\right)\right] = \alpha + \sum_j \beta_j x_{ij},$$

with $x_{ij}$ as the $j$th covariate for person $i$. Measured covariate values from both affected and unaffected individuals were required. An individual's PS, the predicted probability of being affected given the set of covariates, is then calculated as the estimate of

$$\text{PS}_i = \mathrm{e}^{\alpha + \sum_j \beta_j x_{ij}} \Big/ \left(1 + \mathrm{e}^{\alpha + \sum_j \beta_j x_{ij}}\right).$$

### Linkage analysis

Covariate-based affected relative pair linkage analysis using single point IBD probabilities and a general conditional logistic model was performed as implemented in GENIBD and LODPAL of S.A.G.E. v4.4.[8,9,21] In LODPAL, all affected relative pairs are treated as independent, and a single covariate value was calculated for each affected relative pair as the sum of the covariate values for the two affected relatives in the pair. A sum was selected to allow for different risks associated with concordant pairs at the high versus low range of covariate values. LOD scores for the linked and unlinked marker were calculated incorporating no covariates, each covariate alone, both covariates, and the PS as a combined covariate effect.

### Power and type I error rate calculation

The likelihood ratio statistic (LRS) was computed as twice the natural log of the likelihood with versus without using marker data.[8] $P$-values for each replicate were computed assuming a null hypothesis LRS distribution of a mixture of $\chi^2$ distributions. For $K$ covariates, this mixture was 50% ($\chi^2_K$): 50% ($\chi^2_{K+1}$), and if $K = 0$, it was a 50:50 mixture of a point mass at 0 and $\chi^2_1$.[8,9,11,12] Using the appropriate LRS distribution, the proportion of $P$-values among 10 000 replicates that was less than the specified nominal significance value was computed. The power was calculated as this proportion using the linked marker results. The type I error rate was calculated as this proportion using the unlinked marker results.

### Comparison of power across covariate-genetic models

A total of 60 different genetic two-covariate models were simulated. The parameter values for the models are presented in Tables 1a and 1b. Each model contains one binary and one quantitative covariate, and is coupled with three genetic inheritance models (recessive, dominant, and codominant) under two possible penetrances. A baseline model (model0) for covariate functions was selected, and covariate effects were systematically varied (as described by the model features in Table 1b).

### Comparison of the type I error rates by number of covariates included in the analysis for different null hypotheses

To test whether including covariates increased the empirical type I error rate, the results for the unlinked genetic

**Table 1a** Genetic and parameters for two-covariate simulation models

| | *Genetic model* | $p_b(11)$ | $p_b(12)$ | $p_b(22)$ | $q$ |
|---|---|---|---|---|---|
| Penetrance I | Recessive | 0.0 | 0.0 | 0.7 | 0.020 |
| | Dominant | 0.0 | 0.3 | 0.3 | 0.001 |
| | Codominant | 0.0 | 0.1 | 0.3 | 0.001 |
| Penetrance II | Recessive | 0.0 | 0.0 | 0.9 | 0.020 |
| | Dominant | 0.0 | 0.5 | 0.5 | 0.001 |
| | Codominant | 0.0 | 0.3 | 0.5 | 0.001 |

Three genetic models from two possible penetrances (I and II) were simulated for each covariate model.

Abbreviations: $p_b(G)$, baseline disease penetrance for genotype G (11,12, or 22).

**Table 1b** Genetic and covariate parameters for two-covariate simulation models

| | | | | *B* | | | *Q* | | |
|---|---|---|---|---|---|---|---|---|---|
| *Genetic model* | *Model features* | *Covariate model* | $P_B$ | $C_B(11)$ | $C_B(12)$ | $C_B(22)$ | $C_Q(11)$ | $C_Q(12)$ | $C_Q(22)$ |
| Penetrance II | Baseline model | model0 | 0.1 | 0.1 | 0.1 | 0.1 | 0.01 | 0.01 | 0.01 |
| Penetrance I | Low $C_B(G)$, $C_Q(G)$ | model10 | 0.1 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 |
| | Low penetrance | model4 | 0.1 | 0.2 | 0.2 | 0.2 | 0.01 | 0.01 | 0.01 |
| | | model5 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| | | model6 | 0.1 | 0.2 | 0.2 | 0.2 | 0.1 | 0.1 | 0.1 |
| Penetrance II | High penetrance | model1 | 0.1 | 0.2 | 0.2 | 0.2 | 0.01 | 0.01 | 0.01 |
| | | model2 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| | | model3 | 0.1 | 0.2 | 0.2 | 0.2 | 0.1 | 0.1 | 0.1 |
| | Vary binary exposure | bin(0.3) | 0.1 | 0.3 | 0.3 | 0.3 | 0.1 | 0.1 | 0.1 |
| | | bin(0.5) | 0.1 | 0.5 | 0.5 | 0.5 | 0.1 | 0.1 | 0.1 |
| | | bin(0.7) | 0.1 | 0.7 | 0.7 | 0.7 | 0.1 | 0.1 | 0.1 |
| | | bin(0.9) | 0.1 | 0.9 | 0.9 | 0.9 | 0.1 | 0.1 | 0.1 |
| | Increase $P_B$ | model7 | 0.2 | 0.1 | 0.1 | 0.1 | 0.03 | 0.03 | 0.03 |
| | Vary quantitative exposure | quant(0.03) | 0.1 | 0.1 | 0.1 | 0.1 | 0.03 | 0.03 | 0.03 |
| | | quant(0.05) | 0.1 | 0.1 | 0.1 | 0.1 | 0.05 | 0.05 | 0.05 |
| | | quant(0.07) | 0.1 | 0.1 | 0.1 | 0.1 | 0.07 | 0.07 | 0.07 |
| | | quant(0.2) | 0.1 | 0.1 | 0.1 | 0.1 | 0.2 | 0.2 | 0.2 |
| | | quant(0.3) | 0.1 | 0.1 | 0.1 | 0.1 | 0.3 | 0.3 | 0.3 |
| | | quant(0.4) | 0.1 | 0.1 | 0.1 | 0.1 | 0.4 | 0.4 | 0.4 |
| | | quant(0.5) | 0.1 | 0.1 | 0.1 | 0.1 | 0.5 | 0.5 | 0.5 |
| | Interaction | model8 | 0.1 | 0.00 | 0.25 | 0.50 | 0.03 | 0.03 | 0.03 |
| | | model9 | 0.1 | 0.05 | 0.10 | 0.15 | 0.05 | 0.05 | 0.05 |

Three genetic models (from two possible penetrances) were simulated for each covariate model. A baseline covariate model was determined, and varying covariate models were considered with the features listed. For example, the baseline model consisted of model0_recessive, model0_dominant, model0_codominant, with the genetic parameters given by Penetrance II.

Abbreviations: B, binary covariate; Q, quantitative covariate; $P_B$, probability of binary covariate exposure; $C_B$(G),coefficient of genotype-dependent exposure on penetrance for the binary covariate B; $C_Q$(G), coefficient of genotype-dependent coefficient on penetrance for the quantitative covariate Q; G, genotype (11, 12, or 22).

marker locus were compared among the 60 models based on the null hypothesis of no linkage to a genetic locus, but with covariate effects (H1). The type I error rates for the 60 models were averaged separately for each number of covariates included in the linkage analysis: 0, 1 (including the binary covariate, quantitative covariate, and the PS), or 2.

Three additional null hypotheses testing for linkage based on possible biological disease mechanisms were also examined to assess type I error rates:
(H2) a trait with no genetic effect, but with covariate effects
(H3) no linkage to a genetic locus without covariate effects
(H4) a trait with no genetic effect and without covariate effects.

Note that a trait having no genetic effect implies that both the markers simulated at $\theta = 0.01$ and $\theta = 0.5$ should truly be unlinked to the disease. For comparison, three of the original models (model1_dom, model1_codom, and model1_rec) were used to represent the H1 hypothesis. Variations of these models were designed to address the other null hypotheses. A trait with no genetic effect was modeled by setting the baseline penetrances equal across genotypes, using four different levels (0.05, 0.10, 0.20, 0.50), with disease allele frequency of 1 and 0.1% for a total of eight models. The situation of no covariate effect was modeled by setting the covariate coefficients, $C_B(G)$ and $C_Q(G)$, to 0. The type I error rates were calculated for each model from the analysis of the unlinked marker. The rates were then averaged for each number of covariates included in the linkage analysis (0, 1, or 2) over the models appropriate for each null hypothesis.

### Comparison of type I error rates and power with increasing sample size

Twelve models were simulated with increasing samples sizes, from 200 to 500 and 1000 nuclear families ascertained per sample. The three genetic models for model4, model5, model6, and quant0.3 were selected to represent cases with low and moderate baseline power when no covariates were analyzed.

## Results

### Comparison of power among covariate-genetic models

The effect on power of including covariates relative to the analysis with no covariates was consistent across the genetic models and across the nominal significance levels. For a comparison of the covariate-genetic models according to their features, the power at a nominal significance threshold for a representative set according to the covariate analyzed is presented in Table 2. The results suggest that, compared to analyzing no covariates, including any single binary or quantitative covariate may increase or decrease the power to detect linkage, while including a PS consistently increased power more than including either and usually better than including both covariates. Scenarios where the genetic effect was stronger (higher penetrances) or the covariate effect was reduced (lower exposure due to lower $P_B$, or smaller coefficients, $C_B(G)$, $C_Q(G)$, even with low penetrance) lead to higher power. For the limited models where the covariate effect was genotype dependent (interaction), the power with the PS was the highest in three of the six models, and was always higher than analyzing no covariates.

In a comparison of all models according to the covariate analyzed, 68% (41) of the 60 non-interaction models resulted in the PS having the highest power, or matching the highest power $\pm 0.02$. Of that 68%, 35 models resulted in the PS having the highest power, and six models resulted in the PS matching the power of the best method $\pm 0.02$ (no covariate, the binary or quantitative covariate individually, or all covariates). Of the remaining 33% (19 models), the power was $<0.1$ for all of the methods. Consequently, when the maximum power for any of the methods was at least 0.10, including the PS as the single covariate always resulted in the highest power or matched the highest power $\pm 0.02$. Additionally, the estimated covariate effect parameter, $\delta$, from the conditional logistic regression model, was negative for the PS analysis and for the covariates that increased the power when included in the analysis (results not shown).

### Comparison of the type I error rates by number of covariates included in the analysis

The averaged type I error rates for the 60 non-interaction models are presented in Figure 1. One-covariate analyses included binary covariate only, quantitative covariate only, and PS. Increasing the number of covariates analyzed increases the type I error rate by approximately 20% (from 0.051 for no covariates, to 0.059 for one covariate, and to 0.071 for two covariates at a nominal significance level of 0.05). The difference between the nominal and observed rates is greatest at the lower significance levels. As a result, not only is the power often lower when all covariates are included in the analysis as compared to when the PS is used, but an additional penalty results from the inflated type I error rate with both covariates. The type I error rates consistently increased as the included number of covariates was increased, regardless of the null hypothesis assumed (results not shown).

### Comparison of type I error rates and power with increasing sample size

The averaged type I error rates for each type of covariate analyzed under the null hypothesis of no linkage with covariate effects (H1) are presented in Figure 2. The increase in the empirical type I error rates with increasing number of covariates becomes smaller with increasing sample size, suggesting an asymptotically correct null hypothesis LRS distribution, but using this distribution for small sample sizes (below 1000 families) may not be appropriate.

With increasing sample size, the power for these models also increased as expected. The PS resulted in either the highest or matched the highest power $\pm 0.01$ when the baseline power increased above 0.09. As the relative performance of the PS improved with each increase in sample size, the pattern of the power results (which covariate inclusion(s) corresponded to the highest power) also changed with increasing sample size. The power results for one illustrative model for the three sample sizes are presented in Figure 3. This pattern change was not observed when the baseline power dropped to $<0.10$ using stricter nominal significance levels (results not shown). Of the twelve models considered, only the quant0.3 models never reached a power of greater than 0.09 at any sample size.

## Discussion

These extensive simulations compared the number and type of covariates analyzed in a linkage test of a binary trait with covariate effects. Results suggest that including covariates directly in a conditional logistic model can often increase the power to detect linkage compared to including no covariates. Using a PS estimated from multiple covariates can outperform an analysis with multiple covariates, with no covariates, and at the very least will not reduce power even if a covariate which reduces power when analyzed individually is included in its calculation.

This power gain for including the PS is most dramatic when there is moderate power to detect linkage (between 20 and 60%) without covariates. When the baseline power is low (ie $<0.10$) including the PS does not necessarily provide the greatest power. With virtually no power to detect a genetic effect, the pattern of power results is similar to that of type I error, suggesting that the performance of the analytical methods is predominately determined by the type I error behavior. When the sample size increased so that there is some power to detect a genetic effect, the superior performance of the PS over the

**Table 2** Trends in power comparing the various covariate-genetic models

| | | | *Nominal* | *Power for the covariate(s) analyzed* | | | | | *PS* | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Comparisons* | | *Model* | *Threshold* | *None* | *Binary* | *Quant* | *PS* | *Both* | *highest* | *Trends* |
| Genetic models | Codom | model0 | 0.000049 | 0.128 | 0.375 | 0.105 | 0.421 | 0.353 | x | Power of rec/dom≫ codom |
| | Dom | model0 | | 0.877 | 0.978 | 0.846 | 0.983 | 0.970 | x | |
| | Rec | model0 | | 0.947 | 0.999 | 0.943 | 0.999 | 0.994 | x | |
| Penetrances: I (lower) vs II (higher) | I (L) | model4_rec | 0.05 | 0.470 | 0.842 | 0.376 | 0.853 | 0.815 | x | Higher penetrance, higher power |
| | II (H) | model1_rec | | 0.762 | 0.980 | 0.668 | 0.982 | 0.972 | x | |
| | I (L) | model5_dom | | 0.083 | 0.079 | 0.089 | 0.087 | 0.094 | x | |
| | II (H) | model2_dom | | 0.198 | 0.151 | 0.255 | 0.265 | 0.237 | | |
| | I (L) | model6_dom | | 0.070 | 0.068 | 0.079 | 0.073 | 0.082 | | |
| | II (H) | model3_dom | | 0.142 | 0.127 | 0.152 | 0.175 | 0.164 | x | |
| Penetrance I (low) [$C_B(G)=C_Q(G)=C$] | C = 0.1 | model5_codom | 0.05 | 0.084 | 0.080 | 0.090 | 0.089 | 0.099 | | Lower C, higher power |
| | C = 0.03 | model10_codom | | 0.155 | 0.140 | 0.251 | 0.256 | 0.234 | x | |
| Vary binary exposure coefficient [$C_B(G)=C$] | C = 0.1 | model0_dom | 0.05 | 1.000 | 1.000 | 0.998 | 1.000 | 0.996 | x | Lower C, higher power |
| | C = 0.3 | bin(0.3)_dom | | 0.453 | 0.811 | 0.352 | 0.817 | 0.765 | x | |
| | C = 0.5 | bin(0.5)_dom | | 0.165 | 0.407 | 0.128 | 0.413 | 0.362 | x | |
| | C = 0.7 | bin(0.7)_dom | | 0.098 | 0.223 | 0.089 | 0.222 | 0.203 | x | |
| | C = 0.9 | bin(0.9)_dom | | 0.075 | 0.148 | 0.071 | 0.151 | 0.138 | x | |
| Vary quantitative exposure coefficient [$C_Q(G)=C$] | C = 0.01 | model0_rec | 0.05 | 1.000 | 1.000 | 0.999 | 1.000 | 0.996 | x | Lower C, higher power |
| | C = 0.03 | quant(0.03)_rec | | 0.890 | 0.938 | 0.927 | 0.983 | 0.973 | x | |
| | C = 0.05 | quant(0.05)_rec | | 0.545 | 0.532 | 0.693 | 0.761 | 0.731 | x | |
| | C = 0.07 | quant(0.07)_rec | | 0.325 | 0.278 | 0.462 | 0.488 | 0.454 | x | |
| | C = 0.1 | model2_rec | | 0.172 | 0.138 | 0.238 | 0.238 | 0.221 | x | |
| | C = 0.2 | quant(0.2)_rec | | 0.075 | 0.075 | 0.085 | 0.081 | 0.091 | | |
| | C = 0.3 | quant(0.3)_rec | | 0.057 | 0.062 | 0.063 | 0.064 | 0.072 | | |
| | C = 0.4 | quant(0.4)_rec | | 0.055 | 0.065 | 0.062 | 0.058 | 0.074 | | |
| | C = 0.5 | quant(0.5)_rec | | 0.048 | 0.062 | 0.061 | 0.058 | 0.071 | | |
| Increase $P_B$ | $P_B=0.1$ | quant(0.03)_rec | 0.05 | 0.890 | 0.938 | 0.927 | 0.983 | 0.973 | x | Lower $P_B$, higher power |
| | $P_B=0.2$ | model7_rec | | 0.501 | 0.632 | 0.493 | 0.701 | 0.674 | x | |
| Interaction (binary covariate) | Without | model10_codom | 0.05 | 0.155 | 0.140 | 0.251 | 0.256 | 0.234 | x | Power with PS higher than with none |
| | With | model8_codom | | 0.290 | 0.219 | 0.388 | 0.389 | 0.351 | x | |
| | With | model8_dom | | 0.435 | 0.361 | 0.593 | 0.589 | 0.566 | | |
| | With | model8_rec | | 0.059 | 0.065 | 0.063 | 0.063 | 0.072 | | |
| Interaction (binary covariate) | Without | quant(0.05)_dom | 0.05 | 0.625 | 0.596 | 0.717 | 0.792 | 0.739 | x | Power with PS higher than with none |
| | With | model9_codom | | 0.275 | 0.220 | 0.380 | 0.381 | 0.350 | x | |
| | With | model9_dom | | 0.446 | 0.370 | 0.595 | 0.606 | 0.571 | x | |
| | With | model9_rec | | 0.059 | 0.063 | 0.066 | 0.063 | 0.073 | | |

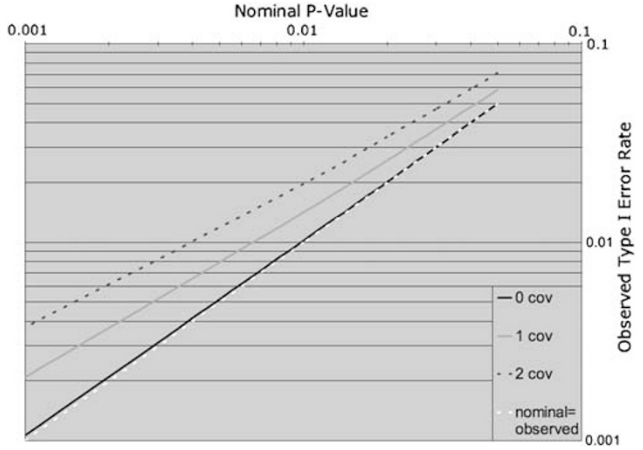

**Figure 1** Observed type I error rates for linkage analysis averaged over all models incorporating 0, 1, or 2 covariates (total of 60 models).

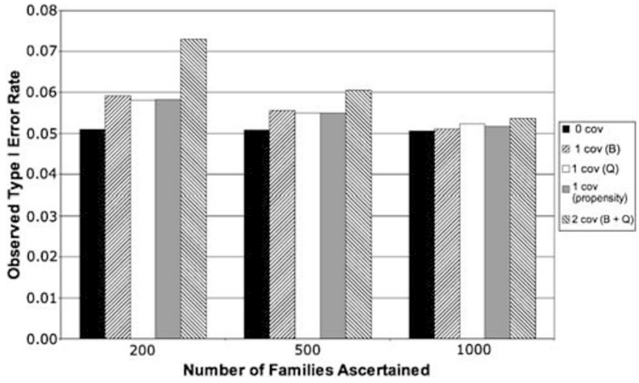

**Figure 2** Observed type I error rates (averaged over 12 models) for the 0.05 nominal significance level according to sample size.

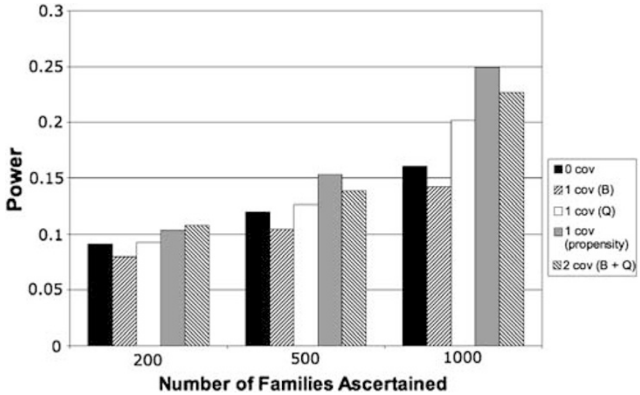

**Figure 3** Power at the nominal significance level of 0.05 for model6_rec by the number of families ascertained.

inclusion of the covariates individually or simultaneously is then observed.

Maximum power was seen with the lowest covariate effect (smallest covariate coefficient in the 'bin' and 'quant' covariate model comparisons, and lower probability of exposure, $P_B$), which seems counter-intuitive. However, with strong covariate effects, disease will be primarily determined by the covariate rather than the genetic effect. The power to detect the genetic effect will decrease, as the ascertained disease population will be heterogeneous with subsets of disease caused by genetic and by nongenetic factors, and the limits of linkage analysis hinge on its ability to detect genetic effects among these various factors.

Power was also reflected in part by the estimated covariate parameter, $\delta$. Using a sum as a pair-specific covariate, situations when the $\delta$ parameter was more negative for the PS covariate had more power. A negative $\delta$ suggests that for each unit decrease in the PS sum ('$x$') for the pair, there was a subsequent increase in the genetic risk, $\lambda = e^{\beta+\delta x}$. This can be expected because affected individuals with low sums of PS are more likely to both have low PS, and so more likely to both have a genetic cause for disease other than through the covariates considered.

An interesting and important finding was the inflation of the type I error rate with increasing number of covariates analyzed. This suggests that for small samples the null hypothesis distribution for the likelihood ratio statistic including $K>0$ covariates is not accurately approximated by a 50:50 mixture of $(\chi^2_K)$ and $(\chi^2_{K+1})$ distributions. Moreover, the larger type I error rate for including both covariates implies that the real power gain for the PS may be larger than the one presently observed. Increasing the sample size reduced this inflation, suggesting that although the distribution of the likelihood ratio statistic may be asymptotically correct, it does not approach its asymptotic distribution with a sample size of 200 families. Consequently, permutation tests or a correction factor is needed to address this issue when the sample size is less than approximately 1000 families.

These simulation results have demonstrated that the use of the PS as a covariate in linkage analysis is promising. However, the strength of the conclusions is limited to the assumptions made regarding the nature of the modeled traits and covariate effects. One major limitation is the simulation of fully observed covariate effects that are independent of each other and familial relationships, and of limited models with gene by covariate interaction. Covariates, such as smoking history/exposure, may have familial correlations or cohort effects, and may be genotype dependent. Thus the true impact of using a PS to increase power may not be as great as these simulations have shown when such correlations and interactions are present. However, because the method of analysis is conditional on families and on covariates, the above correlations do not in principle affect the validity of using a PS. The presence of interactions also adds an additional layer of complexity by limiting the ability to distinguish disease caused by genetic effects and by covariate effects. As such interactions are often not known a *priori*, and including a PS resulted in increased power over including no covariates in the limited models tested, using a PS to increase power will still be valuable although the consistency of its relative performance will need further investigation. Also, because the PS is estimated from fully observed covariates, it is not yet known how missing covariate data in calculating the PS will affect the power of the linkage results or will bias the linkage results. Missing covariate data will impact all analysis methods, and methods that can efficiently estimate the PS with missing covariate data[22] will be valuable in this context.

The second limitation is the requirement of covariate data from both affected and unaffected individuals for calculating the PS. Although covariate data from unaffected individuals are not used in the linkage test, the PS provides valuable information in a linkage analysis, and thus the importance of collecting data on both affected and unaffected individuals needs to be emphasized. The proper ratio of unaffected to affected individuals required to provide a valid estimate of a PS is unclear, but it can be expected that similar to case–control analyses at least one unaffected individual per family is needed. However, many linkage studies of complex traits limit analyses to affected relative pairs, so extensive covariate data may not exist on unaffected individuals and the definition of the PS considered here is not applicable. Moreover, the PS cannot incorporate covariates that are specific to the disease, such as those related to disease severity, which can often be used as classification phenotypes to reduce heterogeneity. If disease is composed of subphenotypes, it may still be possible to analyze only affected data by considering disease specific covariates in calculating a PS for a given subphenotype. Re-classification of affection status can be repeated for the other subtype(s) to identify subtype-specific loci. This approach can also be used to allow for multiple linked loci and potential gene-gene interactions, especially when a linked locus or a gene has been identified and can be used as a covariate. The usefulness of the PS with such covariates still needs to be tested.

The third limitation results because the PS is an estimated variable calculated from a regression on multiple covariates, which is then used to represent a covariate value in the calculation in the likelihood. The degrees of freedom for including a single estimated PS covariate may not be truly one. However, theoretical studies on the application of the PS suggest that paradoxically the use of an estimated value of the PS can provide a more precise balance between the affected and unaffected than the use of the theoretical true value of a PS.[23,24] This, in turn, suggests that treating the value of the PS as a real covariate as opposed to an estimate may produce a more conservative variance for the likelihood ratio statistic.

The fourth limitation is our focus on using the sum as the pair-specific covariate in Olson's conditional logistic regression model,[8] which raises two issues. First, the sum is just one function to express the combined risk from the covariates at which a pair is exposed. Other functions need to be considered with the criterion that they preserve the homogeneity between the two members of the pair within each of the function's levels. For example, using the absolute difference as a function does not allow for risk differences between pairs with both low or high covariate values, which may be important for a covariate like smoking where concordant smokers may have a different genetic risk than concordant nonsmokers, and where covariate effect is independent of genotype. However, with gene by covariate interactions and scenarios where it may be important to distinguish only between concordant and discordant pairs, the use of a difference can be more powerful. Thus, deciding upon an appropriate function for analysis should be highly dictated by the underlying disease model. Second, including a covariate in this conditional logistic model adds two parameters in an unconstrained model, and one parameter in a constrained model. Although the constraint of the one-parameter model may not necessarily be correct under alternative hypotheses, it is correct if there is no linkage conditional on the covariate and so the constraint does not, by itself, invalidate tests for the null hypothesis. However, other possible misspecifications of the model with covariates may result in an increase in the type I error when deviations from the true model are severe. This is a limitation for any model using covariates, as with other parametric models, and is not in principle a limitation of using the PS as a covariate.

The fifth limitation is the simulation of a single locus trait model and use of single point analyses. Although this model may be sufficient to detect linkage for Mendelian diseases, complex diseases will likely consist of multiple loci and will require more powerful multipoint linkage analysis methods. It is possible to simulate more complicated genetic trait models, however, interpretation of results will be dependent on the assumptions of the simulation design, which may not be realistic. Consequently, using simplistic simulation models allows for a more transparent comparison of methodology varying only the number and types of covariates included in the analysis. Additionally, including multiple covariates using the conditional logistic model for multipoint linkage analysis has been shown to increase power to detect linkage in complex diseases, such as for prostrate cancer and Alzheimer's disease.[11,12] Thus, using a PS covariate can also be expected to help when it is also applied to the analysis of real complex diseases.

In summary, with the true nature and number of covariates affecting a genetic trait unknown, using a PS to incorporate multiple covariates into the linkage test is appealing with its increased power over incorporating no or individual covariates, and as a single covariate is subject to lower inflated type I error rates in small sample sizes.

However, this is its first application to covariate-based linkage analysis, and more extensive investigation, especially for the interaction models, is needed. Future plans include studies that will: (1) examine the effect of the PS in models with more than two covariate effects, with gene by covariate interactions, and with varying functions of the covariate pair values (such as with differences, and both sums and differences), (2) determine a correction factor or permutation test to calibrate the power for the inflated type I error when samples sizes are less than 1000 families, (3) further our understanding of the theoretical properties of the PS when applied in the linkage analysis framework, and (4) apply the PS to analyses of real linkage data sets with covariate data available only for affected individuals (though consisting of disease subtypes) and available for both affected and unaffected individuals.

## Acknowledgements

*The results of this paper were obtained by using the program package S.A.G.E., which is supported by a US Public Health Service Resource Grant (RR03655) from the National Center for Research Resources. YYS is partially supported by an American Cancer Society Grant, IRG-58-005-40 and by the Department of Epidemiology, JHBPSH. BQD is supported by a NHGRI pre-doctoral IRTA fellowship. The authors would like to thank the late Jane Olson without whom much of this work would not have been possible, and the two anonymous reviewers for their critical insights.*

## References

1 Greenwood CM, Bull SB: Incorporation of covariates into genome scanning using sib-pair analysis in bipolar affective disorder. *Genet Epidemiol* 1997; **14**: 635–640.

2 Greenwood CM, Bull SB: Analysis of affected sib pairs, with covariates – with and without constraints. *Am J Hum Genet* 1999; **64**: 871–885.

3 Rice JP, Rochberg N, Neuman RJ *et al*: Covariates in linkage analysis. *Genet Epidemiol* 1999; **17** (Suppl 1): S691–S695.

4 Gauderman WJ, Siegmund KD: Gene-environment interaction and affected sib pair linkage analysis. *Hum Hered* 2001; **52**: 34–46.

5 Devlin B, Jones BL, Bacanu SA, Roeder K: Mixture models for linkage analysis of affected sibling pairs and covariates. *Genet Epidemiol* 2002; **22**: 52–65.

6 Shete S, Amos CI, Hwang SJ, Strong LC: Individual-specific liability groups in genetic linkage, with applications to kindreds with Li-Fraumeni syndrome. *Am J Hum Genet* 2002; **70**: 813–817.

7 Mirea L, Briollais L, Bull S: Tests for covariate-associated heterogeneity in IBD allele sharing of affected relatives. *Genet Epidemiol* 2004; **26**: 44–60.

8 Olson JM: A general conditional-logistic model for affected-relative-pair linkage studies. *Am J Hum Genet* 1999; **65**: 1760–1769.

9 Goddard KA, Witte JS, Suarez BK, Catalona WJ, Olson JM: Model-free linkage analysis with covariates confirms linkage of prostate cancer to chromosomes 1 and 4. *Am J Hum Genet* 2001; **68**: 1197–1206.

10 Holmans PA: Asymptotic properties of affected-sib-pair linkage analysis. *Am J Hum Genet* 1993; **52**: 362–374.

11 Olson JM, Goddard KA, Dudek DM: The amyloid precursor protein locus and very-late-onset Alzheimer disease. *Am J Hum Genet* 2001; **69**: 895–899.

12 Olson JM, Goddard KA, Dudek DM: A second locus for very-late-onset Alzheimer disease: a genome scan reveals linkage to 20p and epistasis between 20p and the amyloid precursor protein region. *Am J Hum Genet* 2002; **71**: 154–161.

13 Olson JM, Song Y, Dudek DM *et al*: A genome screen of systemic lupus erythematosus using affected-relative-pair linkage analysis with covariates demonstrates genetic heterogeneity. *Genes Immun* 2002; **3** (Suppl 1): S5–S12.

14 Hill SY, Shen S, Zezza N, Hoffman EK, Perlin M, Allan W: A genome wide search for alcoholism susceptibility genes. *Am J Med Genet* 2004; **128B**: 102–113.

15 Shibamura H, Olson JM, van Vlijmen-Van Keulen C *et al*: Genome scan for familial abdominal aortic aneurysm using sex and family history as covariates suggests genetic heterogeneity and identifies linkage to chromosome 19q13. *Circulation* 2004; **109**: 2103–2108.

16 Rosenbaum PR, Rubin DB: The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.

17 D'Agostino Jr RB: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998; **17**: 2265–2281.

18 Joffe MM, Rosenbaum PR: Invited commentary: propensity scores. *Am J Epidemiol* 1999; **150**: 327–333.

19 Rich SS: Analytic options for asthma genetics. *Clin Exp Allergy* 1998; **28** (Suppl 1): 84–87.

20 Wilson AF, Bailey-Wilson JE, Pugh EW, Sorant AJM: The Genometric Analysis Simulation Program (GASP): a software tool for testing and investigating methods in statistical genetics. *Am J Hum Genet* 1996; **59** (Supp): A193.

21 S.A.G.E. Statistical Analysis for Genetic Epidemiology: Computer program package available from Statistical Solutions Ltd, Cork, Ireland, 2003.

22 D'Agostino Jr RB, Rubin DB: Estimating and using propensity scores with partially missing data. *J Am Statist Assoc* 2000; **95**: 749–759.

23 Rubin DB, Thomas N: Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika* 1992; **79**: 797–809.

24 Rubin DB, Thomas N: Matching using estimated propensity scores: relating theory to practice. *Biometrics* 1996; **52**: 249–264.