# Modifications to the Patient Rule-Induction Method That Utilize Non-Additive Combinations of Genetic and Environmental Effects to Define Partitions That Predict Ischemic Heart Disease

Greg Dyson,[1] Ruth Frikke-Schmidt,[2] Børge G. Nordestgaard,[3,4] Anne Tybjærg-Hansen,[2,4] and Charles F. Sing[1]*

[1] *Department of Human Genetics, University of Michigan, Ann Arbor, Michigan*
[2] *Department of Clinical Biochemistry, Section for Molecular Genetics, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark*
[3] *Department of Clinical Biochemistry, Herlev University Hospital, Herlev, Denmark*
[4] *The Copenhagen City Heart Study, Bispebjerg University Hospital, Copenhagen, Denmark*

This article extends the Patient Rule-Induction Method (PRIM) for modeling cumulative incidence of disease developed by Dyson et al. (Genet Epidemiol 31:515–527) to include the simultaneous consideration of non-additive combinations of predictor variables, a significance test of each combination, an adjustment for multiple testing and a confidence interval for the estimate of the cumulative incidence of disease in each partition. We employ the partitioning algorithm component of the Combinatorial Partitioning Method to construct combinations of predictors, permutation testing to assess the significance of each combination, theoretical arguments for incorporating a multiple testing adjustment and bootstrap resampling to produce the confidence intervals. An illustration of this revised PRIM utilizing a sample of 2,258 European male participants from the Copenhagen City Heart Study is presented that assesses the utility of genetic variants in predicting the presence of ischemic heart disease beyond the established risk factors. *Genet. Epidemiol*. 33 : 317–324, 2009. © 2008 Wiley-Liss, Inc.

**Key words: CPM; PRIM; interaction; non-additivity; IHD**

Contract grant sponsors: Chief Physician Johan Boserup; Lise Boserup's Fund; The Danish Heart Foundation; The Danish Medical Research Council; Ingeborg and Leo Dannin's Grant; The Research Fund at Rigshospitalet, Copenhagen University Hospital; National Institute of Health; Contract grant number: HL072905; Contract grant sponsor: National Institute of General Medical Science; Contract grant number: GM065509.

*Correspondence to: Charles F. Sing, Department of Human Genetics, University of Michigan, 1241 East Catherine Street, 5928 Buhl Building, Ann Arbor, MI 48109-0618. E-mail: csing@umich.edu

Received 6 May 2008; Revised 17 September 2008; Accepted 21 September 2008
Published online 21 November 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20383

## INTRODUCTION

The natural history of ischemic heart disease (IHD) is known to be heterogeneous among individuals within and between gender and racial groups. All individuals within a particular population of inference who develop IHD have not been exposed to the same combination of factors that determine risk. Traditional linear statistical modeling approaches to evaluate the contribution of risk factors ignore this reality. The Patient Rule-Induction Method (PRIM) was first developed by Friedman and Fisher [1999] and incorporated into an approach by Dyson et al. [2007] for ascertaining which values of which measures of the genetic and environmental risk factors predict a discrete disease endpoint in which subset of the sample being studied. This approach yields multiple models, one for each subset of individuals, which is consistent with the assumption that there are multiple etiological pathways responsible for a common disease that has a complex multifactorial etiology in a sample of individuals representative of the population at large [Sing et al., 2003].

The objective of the PRIM is to create mutually exclusive partitions of individuals, defined by terms (selected values of predictor variables), with a higher cumulative incidence than expected under the null distribution. This is achieved through repeated implementations of the peeling and pasting algorithms. Peeling is an iterative process that creates a partition by excluding individuals with particular values of predictor variables, while pasting iteratively amends individuals to the partition, also based upon values of predictor variables, after the peeling stage has been completed [Dyson et al., 2007].

The peeling and pasting algorithms used to create partitions are controlled by two parameters, support and complexity. The support parameter ($\beta$), which is identical for each partition in a PRIM application, is the minimum proportion of unassigned individuals (individuals not already assigned into a partition) that are required to construct a valid partition. The support parameter value for a particular PRIM application is chosen by a grid search. Likelihoods for $\beta$'s, ranging from 0.05 to 0.50 in increments of 0.005 are compared with those for the null model via a likelihood ratio test (LRT) using logistic regression. The null model fits a logistic regression with the intercept as the only predictor of risk. The logistic regression model that considers the set of partitions obtained by a PRIM application using a particular support parameter includes the intercept and a predictor variable that represents the partition class in which an individual was assigned by the corresponding PRIM application. The support parameter which results in the most significant LRT is declared optimal. The complexity parameter, which is pre-selected by the investigator, is the minimum increase in cumulative incidence ($\theta$) in a partition required to add a term to the definition of an existing partition. The statistical significance of the $\theta$ for a partition is tested using permutations created from exchangeable observations. Further details are given in Dyson et al. [2007].

The study of non-additive effects of variations in IHD risk is not typically amenable to traditional linear regression modeling approaches because the relative frequencies of risk factor effects are often correlated and/or there may be rare, or non-existent, multivariable risk factor classes that are characteristic of epidemiological samples representative of the population at large. In an effort to circumvent these possibilities, Nelson et al. [2001] introduced the Combinatorial Partitioning Method (CPM) to ''identify sets of partitions of multi-locus genotypes that predict quantitative trait variability'' in a potentially correlated or sparse space of predictor variables. The partitioning algorithm component of the CPM considers all possible ways of partitioning the values (both additive and non-additive) of a set of categorical variables into $m$ groups and then utilizes an analysis of variance to compare the means of the quantitative trait of the $m$ groups to identify the most statistically significant partition. The CPM was later extended to use a chi-square test for evaluating partitions for categorical traits [Stengård et al., 2006].

This article introduces methodology which (1) combines the features of PRIM and CPM to permit the inclusion of non-additive effects of values of multiple predictor variables in defining peeling and pasting terms and incorporates, (2) permutation testing of the statistical significance of each term used in defining a partition, (3) an adjustment for multiple-testing in establishing the terms that characterize a partition and (4) a confidence interval for the estimate of $\theta$ associated with each partition. These modifications extend the PRIM analysis presented by Dyson et al. [2007] which partitioned the data using only one predictor for each term, tested only the statistical significance of the overall cumulative incidence in a partition, did not correct for multiple testing and lacked a mechanism to compare $\theta$s across partitions. An illustrative application of this modified analytical strategy to test whether combinations of genetic variants improve the ability to predict an IHD event beyond that predicted by the traditional risk factors is presented.

# METHODS

## MODIFICATION OF THE PRIM STRATEGY

### Combining PRIM and CPM

Through the peeling and pasting processes the previously defined PRIM analysis [Dyson et al., 2007] produces mutually exclusive partitions of individuals that are characterized by combinations of values of predictor variables selected one at a time. The incorporation of the CPM partitioning algorithm component facilitates the selection of combinations of values of multiple predictors ($\geq 2$) to define the terms that may be used to characterize the partitions. The algorithm for selection of the value of $\beta$ for a particular application described above is similarly employed in the execution of a PRIM-CPM application.

Table I presents an example of a possible peeling (or pasting) term obtained using the standard PRIM algorithm (a) and the modified PRIM-CPM algorithm (b). The selection of the term defined by value AA of variable SNP2 in Table Ia does not depend on the level of SNP1. The peeling (or pasting) term in Table Ib is defined by four combinations of the levels of two predictor variables considered simultaneously: (SNP1 = CC and SNP2 = AA) or (SNP1 = CC and SNP2 = GG) or (SNP1 = TT and SNP2 = AA) or (SNP1 = TT and SNP2 = GG). This example grouping of genotypes corresponds to the contrast that defines the dominance by dominance non-additive interaction between two locus genotypes in the traditional linear statistical models that have been used in experimental genetics [Cheverud and Routman, 1995]. The partitioning term illustrated in Table Ia is also a possibility that may be constructed by the application of the PRIM-CPM strategy as is any combination of the nine possible pairs of predictor values. The advantage of incorporating the CPM partitioning into the PRIM algorithm is that non-additive effects of two or more variables may be employed to construct terms that may be used to define a partition.

### Determining the statistical significance of each term in each partition

We have modified the PRIM algorithm presented by Dyson et al. [2007] to perform a hypothesis test for each term within each partition at each step in the peeling and pasting processes. This modification eliminates the need for a complexity parameter that serves as a quasi-statistical mechanism for evaluating significance of a term. For any given peeling stage involving $n$ individuals, a support parameter of $\beta$ and a significance level of $\alpha_0$, the PRIM (or PRIM-CPM) algorithm selects the term that produces the subset of individuals ($n_1$) that results in the largest increase in the cumulative incidence of disease, $\theta$, given that $n_1 \geq n^*\beta$. Permuting the observed values of the disease outcome among the $n$ individuals, running the algorithm using the same $\beta$ and returning the resultant $\theta'$ creates one realization of the expectation of $\theta$ under the null distribution. Repeating this procedure $k$ times creates a null distribution for $\theta$ associated with the particular term being

**TABLE I. Standard PRIM (a)[a] and PRIM-CPM (b)[b] examples of a possible peeling or pasting term (meshed squares)**

(a)

| | | SNP2 | | |
|---|---|---|---|---|
| | | AA | AG | GG |
| SNP1 | CC | ▨ | | |
| | CT | ▨ | | |
| | TT | ▨ | | |

(b)

| | | SNP2 | | |
|---|---|---|---|---|
| | | AA | AG | GG |
| SNP1 | CC | ▨ | | ▨ |
| | CT | | | |
| | TT | ▨ | | ▨ |

[a] Illustrates the term (SNP2 = AA), which is not dependent on the levels of SNP1.

[b] Illustrates the term (SNP1 = CC and SNP2 = AA) or (SNP1 = CC and SNP2 = GG) or (SNP1 = TT and SNP2 = AA) or (SNP1 = TT and SNP2 = GG), which is dependent on the levels of both SNP1 and SNP2.

considered. If less than ($100 \times \alpha_0$)% of the $\theta$s obtained from the permutations are greater than the original $\theta$, the term being tested is included in the characterization of the partition and the peeling algorithm then searches for a significant subset of these $n_1$ individuals. If none of the remaining terms significantly increases $\theta$ the peeling process in characterizing this partition is completed and the pasting process is initiated.

Our modification of the pasting process is similar to that adopted for the peeling process. Of the pasting terms (after the completion of the peeling process) that increase $\theta$ (the resultant cumulative incidence for the partition from the completion of the peeling process), the one that results in the largest updated $\theta$ is tested for statistical significance. Permutations of the observed values of the disease outcome among those individuals that have not been assigned to the partition are used to construct the null distribution of the largest $\theta$ anticipated by chance alone for a single pasting term. If less than ($100 \times \alpha_0$)% of the $\theta$s obtained using the permutations are greater than the $\theta$ from this pasting term, then the term is included in the characterization of the partition. Any further pasting terms are similarly tested, using at each step permutations of the disease outcome among the individuals not assigned to a partition at that step in the algorithm. If the inclusion of any of the remaining terms does not significantly increase $\theta$, the pasting process is completed, and the next peeling stage will begin to characterize a new partition using the remaining sample of individuals that have not been assigned to a partition. Since no pasting is attempted unless one or more peeling terms have already defined a partition, the modified PRIM or PRIM-CPM algorithm is terminated when the first peeling term produced from $n$ observations is not statistically significant.

### Multiple testing

Each time a term has the potential to define a partition a hypothesis test is performed using the term testing strategy described above. To lessen the probability of making a Type I error, an experiment-wise correction for multiple testing in the inclusion of multiple terms in the characterization of a partition is required. Since the modified PRIM (or PRIM-CPM) strategy involves sequential hypothesis testing (e.g., a second term is only considered in defining a partition if a first term is statistically significant) a specialized, multiple testing correction is necessary. To achieve this objective we make two assumptions:

(a) $P_{H_0}(T_{i+1}$ is significant$|T_i$ is significant) $\leq P_{H_0}(T_i$ is significant) $\equiv \alpha$ and
(b) $P_{H_0}$(Partition $i+1$ is significant|Partition $i$ is significant) $\leq P_{H_0}$(Partition $i$ is significant),
for sequential tests (of terms within a single partition) $T_i$ and $T_{i+1}$ and partitions $i$ and $i$+1.

Since a second peeling (or pasting) term being considered is conditional on a first peeling (or pasting) term being statistically significant, the probability of making a Type I error on the second peeling (or pasting) term is less than or equal to $P_{H_0}(T_1$ is significant) $\times P_{H_0}(T_2$ is significant$|T_1$ is significant) $= \alpha \times \alpha$, using (a). Note that if $T_1$ is not truly statistically significant and $T_1$ is called significant, a Type I error has already occurred, regardless of the results of $T_2$. The computation of $P_{H_0}(T_2$ is significant$|T_1$ is significant) occurs when $T_1$ is truly statistically significant and called significant and $T_2$ is not truly significant and called significant. The same argument can be made for further tests in the same partition (e.g., $T_3, T_4, \ldots$). Therefore, assuming $p$ peeling (or pasting) terms are produced, the probability of making a Type I error in the peeling (or pasting) stage is less than $\lim_{p\to\infty} \alpha + \alpha^2 + \cdots + \alpha^p = \alpha/(1-\alpha)$. So the probability of making a Type I error for any particular partition, including both peeling and pasting steps, is at most $2\alpha/(1-\alpha)$. Likewise, since the second partition being constructed is conditional on the first one being statistically significant, the probability of making a Type I error if $p$ partitions are produced, using (b), is less than

$$\lim_{p\to\infty} [2\alpha/(1-\alpha) + (2\alpha/(1-\alpha))^2 + \cdots + (2\alpha/(1-\alpha))^p] = 2\alpha/(1-3\alpha)$$

Therefore, if an overall experiment-wise Type I error rate of $\alpha^*$ is desired, each hypothesis test is performed at the $\alpha^*/(2+3\alpha^*)$ level of probability.

### Confidence interval for $\theta$

A confidence interval for the estimated $\theta$ for each partition in the original analysis is produced by bootstrap resampling [Efron and Tibshirani, 1993]. Briefly, we construct a bootstrap sample of the individuals that had the potential to be included in the partition in the original analysis. Therefore individuals already assigned to an earlier partition are not used in creating the bootstrap sample. A PRIM (or PRIM-CPM) application is then performed on this bootstrap sample using the same support parameter and number of peeling and pasting terms as in the original analysis. There is no testing of the significance of each peeling and pasting term as in the original analysis since we want to estimate the distribution of $\theta$ given the number of peeling and pasting terms that were used to characterize the partition in the original analysis. The resultant $\theta'$ for this single partition model associated with a bootstrap sample is a realization from the distribution of the estimated $\theta$ obtained from the original analysis. The 0.025th and 0.975th quantiles of this distribution (generated from 1,000 bootstrap samples) are used to define the 95% confidence interval for the $\theta$ estimated from the original analysis. The same procedure is done for each of the partitions produced by the original analysis.

## ANALYSIS STEPS

Two analyses were carried out on the same example data set, one performing a standard PRIM analysis considering only one variable at a time to define a term and the other using the PRIM-CPM strategy that considers two variables at a time to define a term. In both analyses, the hypothesis testing of each term using the permutation method, the correction for multiple testing described above and the partition confidence interval for $\theta$ were produced. The model building strategy follows two steps in the convention of Dyson et al. [2007], which tests the added predictive value of the genetic variables beyond that obtained using only the traditional risk factors. In the first step only the traditional risk factors are used in the

**TABLE II. Characteristics of male participants in the Copenhagen City Heart Study recruited between 1976 and 1978 and followed until December 31, 1999**

| Covariate | With IHD ($n = 286$) | Without IHD ($n = 1{,}972$) |
|---|---|---|
| Age | | |
| 45–65 | 90 (0.31) | 1,070 (0.54)*** |
| Over 65 | 196 (0.69) | 902 (0.46) |
| Smoking | | |
| No | 89 (0.31) | 717 (0.36) |
| Yes | 197 (0.69) | 1,255 (0.64) |
| Diabetes mellitus | | |
| No | 252 (0.88) | 1,830 (0.93)** |
| Yes | 34 (0.12) | 142 (0.07) |
| Hypertension | | |
| No | 36 (0.13) | 490 (0.25)*** |
| Yes | 250 (0.87) | 1,482 (0.75) |
| Cholesterol | | |
| ≤200 | 45 (0.16) | 380 (0.19) |
| (200, 240) | 101 (0.35) | 752 (0.38) |
| ≥240 | 140 (0.49) | 840 (0.43) |
| HDL-C | | |
| <40 | 68 (0.24) | 374 (0.19) |
| ≥40 | 218 (0.76) | 1,598 (0.81) |
| Triglycerides | | |
| <150 | 118 (0.41) | 962 (0.49)* |
| ≥150 | 168 (0.59) | 1,010 (0.51) |
| BMI | | |
| ≤25 | 89 (0.31) | 737 (0.37) |
| (25,30) | 142 (0.50) | 912 (0.46) |
| ≥30 | 55 (0.19) | 323 (0.17) |
| *APOE* −491A>T (*E560*) | | |
| AA | 204 (0.71) | 1,407 (0.71) |
| AT | 73 (0.26) | 524 (0.27) |
| TT | 9 (0.03) | 41 (0.02) |
| *APOE* −427T>C (*E624*) | | |
| TT | 229 (0.80) | 1,587 (0.80) |
| TC | 55 (0.19) | 368 (0.19) |
| CC | 2 (0.01) | 17 (0.01) |
| *APOE* −219G>T (*E832*) | | |
| GG | 75 (0.26) | 565 (0.29) |
| GT | 150 (0.53) | 991 (0.50) |
| TT | 61 (0.21) | 416 (0.21) |
| *APOE* g.2059T>C (*E3937*) | | |
| TT | 204 (0.71) | 1,391 (0.71) |
| TC | 73 (0.26) | 533 (0.27) |
| CC | 9 (0.03) | 48 (0.02) |
| *APOE* g.2197C>T (*E4075*) | | |
| CC | 246 (0.86) | 1,652 (0.83) |
| CT | 37 (0.13) | 312 (0.16) |
| TT | 3 (0.01) | 8 (0.01) |
| *LPL* g.8756G>A (*LPL9*) | | |
| GG | 279 (0.98) | 1,924 (0.98) |
| GA | 7 (0.02) | 48 (0.02) |
| *LPL* g.16577A>G (*LPL291*) | | |
| AA | 265 (0.93) | 1,866 (0.95) |
| AG | 21 (0.07) | 105 (0.05) |
| GG | | 1 (0.00) |
| *LPL* g.22772C>G (*LPL447*) | | |
| CC | 227 (0.79) | 1,604 (0.81) |
| CG | 56 (0.20) | 346 (0.18) |
| GG | 3 (0.01) | 22 (0.01) |

All exonic sites in *APOE* and *LPL* are named according to human mutation nomenclature [den Dunnen and Antonarakis, 2001]. To correspond with well-established literature names of promoter variants in *APOE,* nucleotide numbering is counted from transcriptional start site. The name in the parentheses is shorthand notation used throughout the article. The combination of E3937 and E4075 SNPs represents the traditional three-allelic [$\varepsilon_2$, $\varepsilon_3$, $\varepsilon_4$] *APOE* polymorphism. IHD, ischemic heart disease.

***Significant at 0.001 level of probability.

**Significant at 0.01 level of probability.

*Significant at 0.05 level of probability.

PRIM (or PRIM-CPM) application to develop predictive models, one for each partition, for incident IHD. In the second step we used PRIM (or PRIM-CPM) applications to evaluate the ability of any of the genotyped single nucleotide polymorphisms (SNPs) to further improve the predictive model of incident IHD in each of the partitions established in the first step.

## EXAMPLE DATA SET

The sample used to illustrate the proposed modified PRIM analysis strategy consists of 2,258 European male participants from the prospective longitudinal general population Copenhagen City Heart Study (CCHS) who enrolled between 1976 and 1978, were IHD free and at least 45 years old at the second follow-up exam (1991–1994), which is treated as the baseline time point for this analysis. These individuals, of whom more than 99% were white and of Danish descent, were followed until December 31, 1999. The study was approved by a Danish ethics committee No. 100.2039/91. The diagnosis of IHD, the definitions and categorizations of all non-genetic predictor variables and the descriptions of the eight SNPs in the *apolipoprotein E* (*APOE*) and *lipoprotein lipase* (*LPL*) genes that were genotyped for this study are presented in the methods given in Dyson et al. [2007].

# RESULTS OF AN APPLICATION TO THE EXAMPLE DATA SET

Table II presents information on each categorical predictor variable used in the two-step PRIM and PRIM-CPM modeling strategies. The relative frequencies of diabetes status, age level, triglyceride level and hypertension status varied significantly between those individuals who experienced an IHD event and those who did not at the 0.05 level. None of the eight SNPs under study had a significant association with the outcome ($\chi^2$ test statistic $P$-value$<$0.05).

A standard PRIM analysis (one variable to define a term in a partition) and the newly developed PRIM-CPM strategy using two variables to define a term to characterize a partition were performed. In each case we utilized the new permutation testing strategy to test the significance of

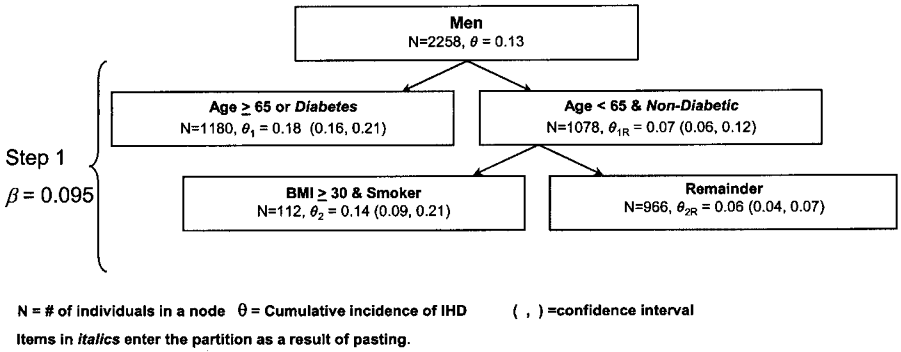
Men
N=2258, θ = 0.13
Age ≥ 65 or Diabetes
N=1180, θ1 = 0.18 (0.16, 0.21)
Age < 65 & Non-Diabetic
N=1078, θ1R = 0.07 (0.06, 0.12)
Step 1
β = 0.095
BMI ≥ 30 & Smoker
N=112, θ2 = 0.14 (0.09, 0.21)
Remainder
N=966, θ2R = 0.06 (0.04, 0.07)
N = # of individuals in a node θ = Cumulative incidence of IHD ( , ) =confidence interval
Items in italics enter the partition as a result of pasting.

Fig. 1. PRIM results using one variable to define each term.

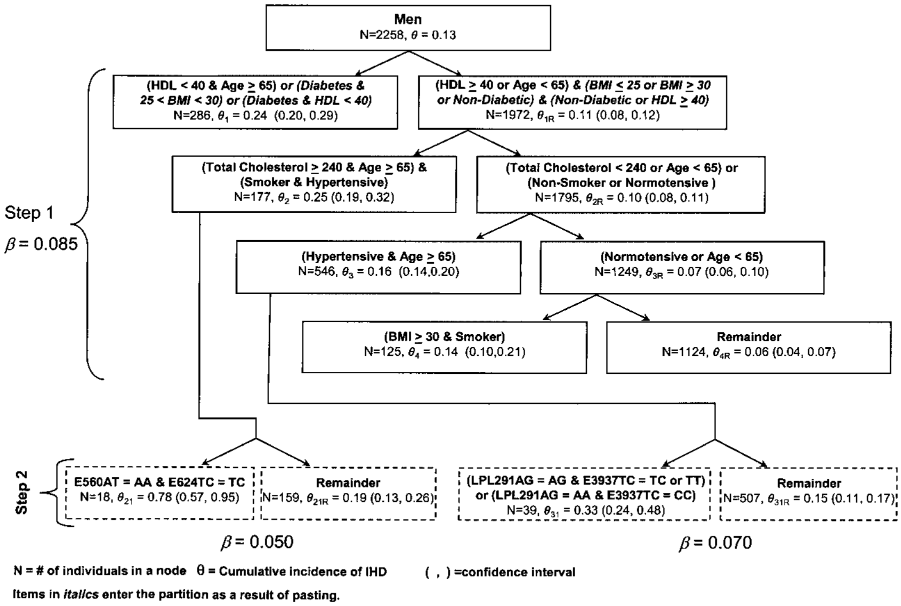
Men
N=2258, θ = 0.13
(HDL < 40 & Age ≥ 65) or (Diabetes & 25 < BMI < 30) or (Diabetes & HDL < 40)
N=286, θ1 = 0.24 (0.20, 0.29)
(HDL ≥ 40 or Age < 65) & (BMI ≤ 25 or BMI ≥ 30 or Non-Diabetic) & (Non-Diabetic or HDL ≥ 40)
N=1972, θ1R = 0.11 (0.08, 0.12)
(Total Cholesterol ≥ 240 & Age ≥ 65) & (Smoker & Hypertensive)
N=177, θ2 = 0.25 (0.19, 0.32)
(Total Cholesterol < 240 or Age < 65) or (Non-Smoker or Normotensive )
N=1795, θ2R = 0.10 (0.08, 0.11)
Step 1
β = 0.085
(Hypertensive & Age ≥ 65)
N=546, θ3 = 0.16 (0.14,0.20)
(Normotensive or Age < 65)
N=1249, θ3R = 0.07 (0.06, 0.10)
(BMI ≥ 30 & Smoker)
N=125, θ4 = 0.14 (0.10,0.21)
Remainder
N=1124, θ4R = 0.06 (0.04, 0.07)
Step 2
E560AT = AA & E624TC = TC
N=18, θ21 = 0.78 (0.57, 0.95)
Remainder
N=159, θ21R = 0.19 (0.13, 0.26)
(LPL291AG = AG & E3937TC = TC or TT) or (LPL291AG = AA & E3937TC = CC)
N=39, θ31 = 0.33 (0.24, 0.48)
Remainder
N=507, θ31R = 0.15 (0.11, 0.17)
β = 0.050
β = 0.070
N = # of individuals in a node θ = Cumulative incidence of IHD ( , ) =confidence interval
Items in italics enter the partition as a result of pasting.

Fig. 2. PRIM-CPM results using two variables to define each term.

a peeling or pasting term, corrected for multiple tests and applied the bootstrap procedure to compute confidence intervals for the estimated values of $\theta$. To ensure an overall Type I error rate of 0.10, we tested every hypothesis at $\alpha = 0.10/(2+3*0.10) = 0.043$, following the rationale given in methods section. Figures 1 and 2 display the results of PRIM using one variable at a time and of PRIM-CPM using two variables at a time to define a term, respectively.

## STANDARD PRIM: ONE VARIABLE TO DEFINE A TERM (FIG. 1)

*Step 1 (traditional risk factors)*: The overall cumulative incidence of an IHD event during 8 years of surveillance for male CCHS participants was 0.13. The $\beta$ parameter chosen by the algorithm described in Dyson et al. [2007] for this PRIM application was 0.095. The first partition included 1,180 individuals who were aged 65 and older or were diabetic ($P_1$, $\theta_1 = 0.18$, confidence interval (CI): (0.16, 0.21)). The second partition consisted of the 112 individuals out of the 1,078 unassigned individuals (non-diabetic individuals less than 65 years old) who had BMI≥30 and smoked ($P_2$, $\theta_2 = 0.14$, CI: (0.09, 0.21)). The remaining 966 individuals make up the remainder group ($P_{2R,}$ $\theta_{2R} = 0.06$, CI: (0.04, 0.07)).

*Step 2 (genetic variations)*: There were no significant genetic effects when individual SNPs were used to define the peeling and pasting terms in any of the three partitions from step 1.

## PRIM-CPM: TWO PREDICTOR VARIABLES TO DEFINE A TERM (FIG. 2)

*Step 1 (traditional risk factors)*: The $\beta$ parameter chosen for this PRIM-CPM application was 0.085. The first partition included 286 individuals who had HDL-C less than 40 and were at least 65 years of age or were diabetic and had a BMI between 25 and 30 or were diabetic and had HDL-C<40 ($P_1$, $\theta_1 = 0.24$, CI: (0.20, 0.29)). The second partition consisted of the 177 individuals out of the 1,972 unassigned individuals who had a total cholesterol greater than or equal to 240 and were 65 years or older and were hypertensive and smokers ($P_2$, $\theta_2 = 0.25$, CI: (0.19, 0.32)). The third partition consisted of the 546 individuals out of the 1,795 unassigned individuals were 65 years of age or older and were hypertensive ($P_3$, $\theta_3 = 0.16$, CI: (0.14, 0.20)). The fourth partition consisted of the 125 individuals out of the 1,249 unassigned individuals who had a BMI equal to or greater than 30 and were smokers ($P_4$, $\theta_4 = 0.14$, CI: (0.10, 0.21)). The 1,124 individuals that make up the remainder group ($P_{4R}$) had $\theta_{4R|}$ of 0.06, CI: (0.04, 0.07).

***Step 2 (genetic variations)***: Further partitioning of each of the five partitions produced in step 1 ($P_1$, $P_2$, $P_3$, $P_4$, and $P_{4R}$ in Fig. 2) was carried out using information on two genetic variations at a time to define a term. The 177 individuals assigned to partition $P_2$ were further partitioned by a PRIM application (using a $\beta$ of 0.050) into two groups using SNPs E560 and E624 to define terms considered in the predictive model. Eighteen individuals who were heterozygous (TC) for E624 and homozygous for the most frequent homozygote (AA) for E560 had a significantly higher cumulative incidence than individuals in all other genotype classes ($P_{21}$, $\theta_{21} = 0.78$ (CI: (0.57, 0.95)) versus 0.19 (CI: (0.13, 0.26))). The 546 individuals assigned to partition $P_3$ were further partitioned by a PRIM application (using a $\beta$ of 0.070) into two groups using SNPs LPL291 and E3937 to define terms considered in the predictive model. Thirty-nine individuals who were in one of the following three genotype combinations: (LPL291 = AG and E3937 = TC), (LPL291 = AA and E3937 = CC), (LPL291 = AG and E3937 = TT) had a significantly higher cumulative incidence than individuals in all other genotype classes ($P_{31}, \theta_{31} = 0.33$ (CI: (0.24, 0.48)) versus 0.14 (CI: (0.11, 0.17))).

## DISCUSSION

The past several decades of human genetic research have generated great enthusiasm for the utility of information about genomic variation for understanding and predicting common diseases that have a complex multifactorial etiology [Dollery, 2007; Guttmacher and Collins, 2005]. The most impressive progress toward this goal has been in the development of high throughput laboratory methods to measure DNA sequence variations [Mardis, 2008]. Research to understand the complex causal relationships between genome sequence variation and emergent clinical endpoints through dynamic metabolic networks is in its infancy [Benfry and Mitchell-Olds, 2008; Loscalzo et al., 2007]. Most practical research focuses on the evaluation of the ability of genomic variations to statistically predict inter-individual variation in intermediate metabolites and clinically defined phenotypes of disease. However such research is often divorced from the reality about the limited impact of genetic variants in clinical practice and the inherent complexity of the biological systems that link genomic variation with variation in risk of disease. The PRIM and PRIM-CPM strategies address three goals of research to develop pragmatic prediction models for complex disease endpoints: (1) evaluation of the added predictive value of genomic information beyond traditional risk factors, (2) modeling that recognizes the inherent etiological heterogeneity among subdivisions of the population at large and (3) identification of non-additive effects of genetic and environmental predictors that take into account the reality of correlated frequencies of predictor values and a sparse space of multivariable combinations that are typical of non-experimental data representative of the human population.

Because of the historical evolution of the incorporation of biological information into the practice of medicine, the primary goal of such statistical research is to determine which of the millions of genomic variations add value to prediction beyond those traditional risk factors and biomarkers that have been accepted by medicine as having utility in the clinic for evaluating risk of disease and predicting progression of disease and response to therapy. The two-step strategy for applying the modified PRIM and PRIM-CPM algorithms acknowledges this reality.

It is widely acknowledged in clinical medicine that the etiology of a complex multifactorial disease is heterogeneous among patients and families. The search for those genomic variations that have clinical utility using traditional regression modeling approaches ignores the possibility that variations in different genes are relevant for determining disease endpoints in different subsets of the population at large. Given this context-dependent reality, the analytical question becomes which combination of predictor variables, in which subset of individuals of the population defines the best prediction model of the endpoint of interest?; not which variables are the best predictors in a model that is assumed to be appropriate for every individual of the population at large. We propose the modified PRIM and PRIM-CPM algorithms as a non-traditional, non-parametric, alternative statistical strategy for building multiple prediction models that acknowledge and reflect the etiological realities of a common human disease.

Three of the four major modifications of the PRIM algorithm that are introduced in this article address hypothesis testing issues that have parallels in the application of traditional parametric approaches: significance testing of each combination of predictor variables potentially used in defining a partition, an adjustment to correct the number of hypothesis tests carried out in selecting the terms in the peeling and pasting processes to characterize a partition and a confidence interval for $\theta$ that allows for comparisons to be made between partitions. Permutation testing of the statistical significance of each term in the building of each partition is expected to result in fewer terms than testing the significance of the estimate of $\theta$ associated with the final partition. Furthermore, the problem of over-determination and sparseness inherent when using the CPM partitioning structure in high dimensions is expected to be lessened by performing a permutation test of the significance of each term considered for inclusion in the prediction model. The expected result is a more parsimonious and robust set of models defined by the PRIM (or PRIM-CPM) application. Although this improvement is tenable when the number of predictor variables is in the hundreds, the strategy quickly exceeds the computational limits when the number of variables is in the millions. To address this limitation, we are currently developing a theoretical, non-permutation based method that tests every term in every partition. This alternative will alleviate the computational limit imposed by permutation testing to allow applications of the proposed partitioning strategy to data sets that include millions of predictor variables (e.g., genome-wide association studies).

While some researchers [Talmud et al., 2002; Wright et al., 2006] advocate lowering the threshold of statistical significance as a way to correct the experiment-wise error for multiple testing, we acknowledge that performing many sequential, correlated hypothesis tests in defining the terms that characterize a partition requires an alternative correction mechanism. The derivation of the multiple testing adjustment approach introduced in this article (applicable regardless of the number of predictor variables and consequently the number of potential terms) requires two assumptions: (1) the probability that a peeling or pasting term at any step in the modeling building process is significant is less than or equal to the probability that the previously incorporated peeling or pasting term is statistically significant and (2) the probability that a completed partition is significant at any step in the modeling building process is less than or equal to the probability that the previously defined partition is statistically significant. Insights into the validity of these assumptions cannot be made from the application to one data set. Only large-scale simulation studies considering a broad range of risk variable-endpoint etiologies to evaluate these assumptions could help resolve the utility of the experiment-wise error rate adjustment that we have proposed here. The inferences about these assumptions that are possible from simulation studies will depend entirely on the domain of possibilities defined by parameter values that are unknown, or unknowable, for a particular population of inference.

The bootstrap approach to producing a confidence interval for $\theta$ estimated for a partition introduced in the modification of the PRIM plays a key role in the comparisons between partitions. Determining if a $\theta$ from a partition is statistically significantly different than a $\theta$ from another partition enables the investigator to evaluate the distribution of $\theta$ among partitions and make practical decisions about the utility of particular partitions in clinical practice.

The analysis of the example data set illustrates the presence of etiological heterogeneity and non-additive influences of predictor variables. Both the analyses of one variable at a time and two variables at a time illustrate the etiological expectation that different variables define terms that predict IHD in different subgroups of the sample. When traditional risk factors are considered one at a time, age, diabetes and smoking contribute to partitioning the sample into three subgroups of individuals (Fig. 1). When risk factors are considered two at a time to define a partition these predictor variables combine non-additively with BMI, total cholesterol, HDL-C and hypertension to define five statistically significant partitions (Fig. 2). Comparing the $P_2$ partition, Fig. 1, established in the analysis of one variable at a time with the $P_4$ partition, Fig. 2, established by the analysis two variables at a time suggests that smoking status combines with BMI to predict IHD in two different contexts. However, the analysis of two variables at a time also illustrates that smoking combines non-additively with total cholesterol, age and hypertension to define a partition ($P_2$, Fig. 2). Consideration of the confidence intervals associated with the estimates of $\theta$ for the five subgroups suggests three risk groups: ($P_1 = 0.24$ and $P_2 = 0.25$), ($P_3 = 0.16$ and $P_4 = 0.14$) and ($P_{4R} = 0.06$).

The expected etiological role that non-additive genetic and environmental effects have in the development of IHD is illustrated by the PRIM-CPM application presented in step 2, Figure 2. Particular two locus genotypic variations have added value in defining partitions only in particular partitions established in the step 1 application of the PRIM-CPM (Fig. 2 and Table III). Two variations in the 5′ of the *APOE* gene identify a subgroup with a significant increase in risk ($\theta_{21} = 0.78$ versus $\theta_{21R} = 0.19$) only in partition 2. Similarly, the two locus genotypes defined by the combination of *LPL* and *APOE* SNPs define a statistically significant increase in risk ($\theta_{31} = 0.33$ versus $\theta_{31R} = 0.15$) only in partition 3. The same combination of two 5′ *APOE* SNPs were selected to define the same partition of individuals when step 2 analysis considered variables one at a time. Neither the LPL291 nor the E937 SNP was selected to define a partition when considered one at a time suggesting non-additivity in their contribution to predicting IHD. Before any substantive conclusions regarding the validity of the inferences from the models defined in this study are drawn, the models should be validated in another independent sample from the same population of inference, which is a relevant issue regardless of the model building strategy employed.

As $\beta$ was defined to be larger than 0.05 in the peeling process, rare combinations of values of one or more predictor variables (including genotypes) can only be used to define terms that characterize partitions in the pasting

**TABLE III. Comparison of significant PRIM-CPM-defined genotype contrasts across step 1 partitions**

| SNPs | Contrast | All partitions | Partition 1 | Partition 2 | Partition 3 | Partition 4 | Remainder |
|---|---|---|---|---|---|---|---|
| E560/E624 | AA/TC | 0.13 (355) | 0.20 (46) | 0.78 (18) | 0.14 (98) | 0.11 (19) | 0.05 (174) |
| | Others | 0.13 (1,903) | 0.25 (240) | 0.20 (159) | 0.16 (448) | 0.15 (106) | 0.06 (950) |
| LPL291/E3937 | (AA/CC,AG/TC,AG/TT) | 0.17 (181) | 0.18 (34) | 0.22 (18) | 0.33 (39) | 0.29 (7) | 0.06 (83) |
| | Others | 0.12 (2,077) | 0.25 (252) | 0.26 (159) | 0.15 (507) | 0.14 (118) | 0.06 (1,041) |
| Overall | – | 0.13 (2,258) | 0.24 (286) | 0.25 (177) | 0.16 (546) | 0.14 (125) | 0.06 (1,124) |

Significant contrasts are bordered.
Cells in table display the cumulative incidence and sample size in parentheses.

process. This is illustrated through the inclusion of the (diabetic and $25<\mathrm{BMI}<30$) and (diabetic and HDL-C $<40$) terms that characterize partition 1 of the PRIM-CPM analysis (Fig. 2). Since in the entire sample only 62 (2.7%) and 70 (3.1%) individuals, respectively, are in these two low relative frequency terms, these terms could only characterize a partition in the pasting process. Although no rare combination of genotypes resulted in characterizing a partition in the step 2 (genetic) analysis in either example presented, the pasting process would be able to identify such a context-dependent effect if it existed within a partition of individuals defined by traditional risk factors in step 1.

*APOE and LPL* are key components in human lipid metabolism, mediating the clearance and modulation of triglyceride-rich lipoproteins [Brunzell and Deeb, 2001; Mahley and Rall, 2001]. Genetic variants in the two genes encoding these proteins have been extensively studied in human populations and influence the inter-individual variation in levels of cholesterol and triglycerides [Brunzell and Deeb, 2001; Davignon et al., 1988; Frikke-Schmidt et al., 2000; Stengård, 2006]. Their contribution as single sites to prediction of IHD in the general population as a whole has, however, been subtle [Frikke-Schmidt et al., 2007; Wittrup et al., 2006]. The example analysis presented here illustrates that the incidence of a common disease having a complex multifactorial etiology may be influenced by gene–gene and/or gene–environment interactions that result in high-risk subgroups of the population defined by different combinations of genetic and environmental factors. We suggest that such genetic/biological interactions will have a higher likelihood of being detected with statistical strategies that consider how many predictive models there are for a particular population of inference rather than which variables should be included in a single prediction model.

In summary, the application of PRIM and PRIM-CPM to a large example data set illustrates the selection of multiple models for an etiologically heterogeneous complex disease endpoint, added value of genetic variation dependent on context defined by traditional risk factors and non-additivity of predictor variables at three levels, between variables in step 1 (both term-level non-additive effects and partition-level non-additivity), between genetic variations in step 2 and between traditional predictor variables in step 1 and genetic variations in step 2. Multiple prediction models that incorporate these realities are expected to improve the utility of genetic variations in the practice of medicine and in the design of studies to understand the role of non-additive effects of the many factors that contribute to the etiology of a disease such as IHD.

## ACKNOWLEDGMENTS

The authors thank Mette Refstrup, Pia T. Petersen and Hanne Damm for expert technical assistance, Kenneth G. Weiss for his dedicated attention to the details of the data management, Richard Merkle and Alex Sadovsky for their assistance in programming and data analysis.

## REFERENCES

Benfry PN, Mitchell-Olds T. 2008. From genotype to phenotype: systems biology meets natural variation. Science 320:495–497.

Brunzell JD, Deeb SS. 2001. Familial lipoprotein lipase deficiency, apoC-II deficiency, and hepatic lipase deficiency. In: Scriver CR, Beaudet AL, Sly S, Valle D, editors. The Metabolic and Molecular Bases of Inherited Disease, 8th edition. New York: McGraw-Hill, p. 2789–2816.

Cheverud JM, Routman EJ. 1995. Epistasis and its contribution to genetic variance-components. Genetics 139:1455–1461.

Davignon J, Gregg RE, Sing CF. 1988. Apolipoprotein E polymorphism and atherosclerosis. Arteriosclerosis 8:1–21.

den Dunnen JT, Antonarakis SE. 2001. Nomenclature for the description of human sequence variations. Hum Genet 109:121–124.

Dollery CT. 2007. Beyond genomics. Clin Pharmacol Ther 82:366–370.

Dyson G, Frikke-Schmidt R, Nordestgaard BG, Tybjærg-Hansen A, Sing CF. 2007. An application of the patient rule-induction method for evaluating the contribution of the Apolipoprotein E and Lipoprotein Lipase genes to predicting ischemic heart disease. Genet Epidemiol 31:515–527.

Efron B, Tibshirani R. 1993. An Introduction to the Bootstrap. New York: Chapman & Hall.

Friedman JH, Fisher NI. 1999. Bump hunting in high dimensional data. Stat Comput 9:123–143.

Frikke-Schmidt R, Nordestgaard BG, Agerholm-Larsen B, Schnohr P, Tybjærg-Hansen A. 2000. Context dependent and invariant associations between lipids lipoproteins and apolipoproteins and apolipoprotein E genotype. J Lipid Res 41:1812–1822.

Frikke-Schmidt R, Sing CF, Nordestgaard BG, Steffensen R, Tybjærg-Hansen A. 2007. Subsets of SNPs define rare genotype classes that predict ischemic heart disease. Hum Genet 120:865–877.

Guttmacher AE, Collins FS. 2005. Realizing the promise of genomics in biomedical research. JAMA 294(11):1399–1402.

Loscalzo J, Kohane I, Barabasi AL. 2007. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Mol Syst Biol 3:124.

Mahley RW, Rall SC. 2001. Type III Hyperlipoproteinemia Dysbetalipoproteinemia: The Role of Apolipoprotein E in Normal and Abnormal Lipoprotein Metabolism. In: Scriver CR, Beaudet AL, Sly S, Valle D, editors. The Metabolic and Molecular Bases of Inherited Disease, 8th edition. New York: McGraw-Hill, p. 2835–2862.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141.

Nelson MR, Kardia SLR, Ferrell RE, Sing CF. 2001. A combinatorial partitioning method to identify multilocus genotypic parti tions that predict quantitative trait variation. Genome Res 11:458–470.

Sing CF, Stengard JH, Kardia SLR. 2003. Genes, environment, and cardiovascular disease. Arterioscl Throm Vas 23:1190–1196.

Stengård JH, Kardia SLR, Hamon SC, Frikke-Schmidt R, Tybjaerg-Hansen A, Salomaa V, Boerwinkle E, Sing CF. 2006. Contribution of regulatory and structural variations in APOE to predicting dyslipidemia. J Lipid Res 47:318–328.

Talmud PJ, Hawe E, Martin S, Olivier M, Miller GJ, Rubin EM, Pennacchio LA, Humphries SE. 2002. Relative contribution of variation within the APOC3/A4/A5 gene cluster in determining plasma triglycerides. Hum Mol Genet 11:3039–3046.

Wittrup HH, Andersen RV, Tybjaerg-Hansen A, Jensen GB, Nordestgaard BG. 2006. Combined analysis of six lipoprotein lipase genetic variants on triglycerides, high-density lipoprotein, and ischemic heart disease: cross-sectional, prospective, and case-control studies from the Copenhagen City Heart Study. J Clin Endocr Metab 91:1438–1445.

Wright WT, Young IS, Nicholls DP, Patterson C, Lyttle K, Graham CA. 2006. SNPs at the APOA5 gene account for the strong association with hypertriglyceridaemia at the APOA5/A4/C3/A1 locus on chromosome 11q23 in the Northern Irish population. Atherosclerosis 185:353–360.