Biometrical Journal **46** (2004) 6, 707–719 DOI: 10.1002/bimj.200310075

A Mixture-of-Genotypes Model for the Distribution of Thermostable Phenol Sulfotransferase Activity

Jungnam Joo[*, 1], **Hongshik Ahn**[**, 1], **Robert R. Delongchamp**[2], **Susan A. Nowell**[3], and **Nicholas P. Lang**[4]

[1] Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11790, USA
[2] Divison of Biometry and Risk Assessment, National Center for Toxicological Research, FDA, Jefferson, AR 72079, USA
[3] Cancer Prevention and Population Sciences, Roswell Park Cancer Institute, Buffalo, NY 14263, USA
[4] Central Arkansas Veterans Healthcare System and Department of Surgery, College of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA

Received 3 September 2003, revised 26 August 2004, accepted 27 September 2004

Summary

A statistical method for parametric density estimation based upon a mixture-of-genotypes model is developed for the thermostable phenol sulfotransferase (SULT1A1) activity which has a putative role in modifying risk for colon and prostate cancer/polyps. The EM algorithm for the general mixture model is modified to accommodate the genetic constraints and is used to estimate genotype frequencies from the distribution of the SULT1A1 phenotype. A parametric bootstrap likelihood ratio test is considered as a testing method for the number of mixing components. The size and power of the test is then investigated and compared with the conventional chi-squared test. The relative risk associated with genotypes defined by this model is also investigated through the generalized linear model. This analysis revealed that a genotype with the highest mean value of SULT1A1 activity has greater impact on cancer risk than others. This result suggests that the phenotype with a higher SULT1A1 activity might be important in studying the association between the cancer risk and SULT1A1 activity.

Key words: Bootstrap; Density estimation; EM algorithm; Genotype; Phenotype.

1 Introduction

Thermostable phenol sulfotransferase (SULT1A1) plays a role in the metabolism of heterocyclic amines (Kaderlik and Kadlubar, 1995). Because of this role, genetic variants of SULT1A1 potentially modify the risk of colon cancer or colon polyps that has been associated with dietary exposures to heterocyclic amines, e.g., overcooked meats. Frame et al. (2000) developed an assay for SULT1A1 and measured SULT1A1 activity phenotype of subjects in a case-control study of colon cancer.

SULT1A1 activity toward 2-naphthol was assessed using platelets collected in blood samples from Arkansas populations (Frame et al., 2000). The case-control study was designed to match controls to sex, age and race of cases. Subjects with lower values than the cut-off are labeled ‘slow phenotype’, and those with higher values than the cut-off are labeled ‘fast phenotype’. The cut-off is selected by a graphical procedure (Jackson et al., 1989). The data contain 121 controlled subjects and 46 colon cancer cases with ages (ranged from 20 to 86), sex, and race (White, Black and Asian).

[*] Present address: Office of Biostatistics Research, NHLBI, Bethesda, MD 20892
[**] Corresponding author: e-mail: hahn@ams.sunysb.edu

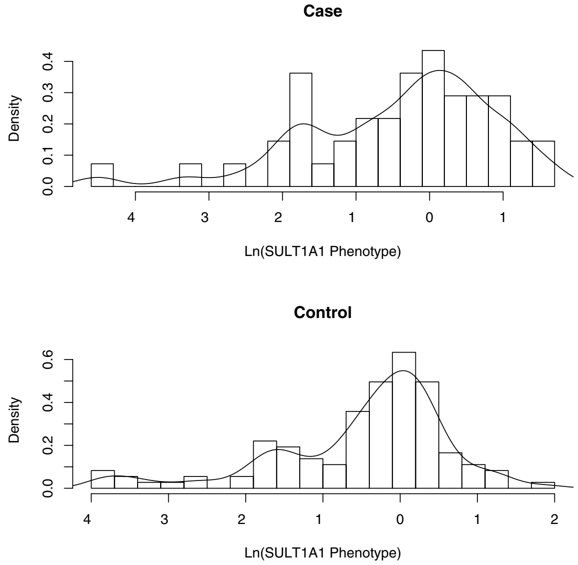

Figure 1 Histograms and kernel density estimates of logarithm of SULT1A1 Phenotype from 46 colon cancer cases and 121 controls.

Figure 1 displays the histograms and kernel density estimates of SULT1A1 activities for the colon cancer case and control. This figure illustrates that the data can be described by three to four prominent components. The probability distribution of SULT1A1 activities in a population is naturally expressed as the mixture of activities associated with each genotype.

The mixture model approach (Elston and Stewart, 1971; Morton and MacLean, 1974; Boyle and Elston, 1979) is a useful tool with wide applicability in the field of genetic epidemiology. EM algorithm is a general approach to maximum likelihood estimation for a mixture distribution (Morton and MacLean, 1974; Dempster et al., 1977; McLachlan and Basford, 1988). Under fairly mild regularity conditions, EM can be shown to converge to a local maximum of the observed data likelihood. Although these conditions do not always hold in practice and the convergence of EM iteration is very slow when the components are not well separated, EM algorithm has been widely used for maximum likelihood estimation for mixture models with good results.

An important question in applying a mixture model is the determination of the number of components. One easy way is to use the likelihood ratio test statistic. It is known that under regularity conditions, this statistic asymptotically follows a chi-squared distribution with degrees of freedom equal to the difference between the number of parameters under the null and alternative hypotheses (Cox and Hinkley, 1974). However, regularity conditions do not hold with the mixture models, since the mixing proportions lie on the boundary of the parameter space under the null hypothesis (Wolfe 1970; Binder, 1978). Goffinet et al. (1992) found the exact limiting distribution in several problems, but the weight parameters were assumed to be known under the alternative hypothesis. In general, however, there is little known beyond the fact that the standard theory does not apply (Böhning, 1998, p. 86).

One successful approach to approximate the null distribution of tests on parameters in the mixture approach is the parametric bootstrap. Aitkin et al. (1981) assessed the null distribution of the likelihood ratio test statistic using a resampling method which is a particular application of the general bootstrap approach (Efron 1979, 1982). McLachlan (1987) developed a bootstrap method for assessing the null distribution of the log likelihood ratio test statistic for the test of a single normal density versus a mixture of two normal densities in the univariate case. Soromenho (1994) compared the performance of five different approaches for testing the number of components in a finite mixture model. He concluded that the bootstrap approach and a procedure based on a stochastic EM procedure yield higher percentages of correct identification of the true model and achieve higher empirical power (Lo et al., 2001). We extend the method of McLachlan (1987) to a mixture with more than two components for an application to the SULT1A1 data in this paper. We investigate this method for determining the number of components. Further, a simulation study for size and power evaluation of the parametric bootstrap likelihood ratio test (LRT) is conducted for the SULT1A1 activities in colon cancer case-control study.

The estimation of the relative risk associated with each genotype in the case-control study can be carried out with the standard logistic regression when the genotype of each subject is known. In our data, a genotype of each subject is not known and therefore inferred from a phenotype measurement through a mixture-of-genotypes model, so it is not clear how to assess the cancer risk. A promising approach to estimating the cancer risk of putative genotypes is to replace an indicator vector for these genotypes in a standard logistic regression model with an expectation for this vector given the observed phenotype (Delongchamp 1993). We apply this method to estimate the cancer risk in this SULT1A1 study.

The main purpose of this paper is to characterize the distribution of SULT1A1 activities in population and to assess the cancer risk associated with these estimated genotypes in the colon cancer case-control study. We also applied proposed methods to the recently obtained data set which contains SULT1A1 activities (phenotypes) in the prostate cancer case and control subjects. SULT1A1 has been implicated in numerous detoxification and bioactivation pathways. However, little is known regarding its endogenous function or its putative role in mediating risk for human environmental disease (Frame et al., 2000). Our methods for phenotyping SULT1A1 activities may help researchers assess a role for this enzyme in disease susceptibility.

The application of proposed methods is not restricted to SULT1A1 data. They are applicable to general case-control studies with a putative marker providing only phenotypic information of a marker but not genotypic information. For example, an assay involving caffeine metabolites, collected in urine samples, indicates the cytochrome P450 CYP1A2 (CYP2A6) phenotype (Butler et al., 1992; Lang et al., 1994; Nowell et al., 2002). We can apply the proposed methods to investigate the association between CYP2A6 activity and colorectal cancer.

2 Material and Methods

2.1 The mixture-of-genotypes model

Suppose that the probability distribution of a given phenotype x_i can be expressed as a mixture of distributions associated with each genotype. If there are K genotypes having relative frequencies, $\tau_1, \dots, \tau_K$, and $f_k(x_i)$ denotes the probability density which results through the expression of a genotype k with parameter θ_k. Then the likelihood for a mixture model with K-components is

$$\prod_{i=1}^{n} \sum_{k=1}^{K} \tau_k f_k(x_i \mid \theta_k) .$$

Since allele frequencies in the studied population determine the mixing proportions, the likelihood is decomposed into that of homozygous and heterozygous subjects using allele frequencies instead of genotype frequencies. Suppose that there are a alleles having relative frequencies $\eta_1, \dots, \eta_a$, and that $g_{ij}(x)$ denotes the probability density for SULT1A1 activity that results through the expression of a genotype composed of alleles i and j. If random segregation of alleles during meiosis largely determines the mixing proportions, observed data have the mixture distribution given as

$$\sum_{i=1}^{a} \sum_{j=1}^{a} \eta_i \eta_j g_{ij}(x) .$$

Since $g_{ij}(x) = g_{ji}(x)$, this distribution can be written as a sum of distinct genotypes

$$\sum_{i=1}^{a} \eta_i^2 g_{ii}(x) + 2 \sum_{i=1}^{a-1} \sum_{j=i+1}^{a} \eta_i \eta_j g_{ij}(x) .$$

The normal density is the most commonly used density for f_k or g_{ij}.

Maximum likelihood estimation for a given mixture model can be carried out using the EM (Expectation-Maximization) algorithm (Dempster et al., 1977). By introducing missing observation of the data $z_i = (z_{i1}, \dots, z_{iK})$ with

$$z_{ik} = \begin{cases} 1 & \text{if } x_i \text{ belongs to group } k, \\ 0 & \text{otherwise} \end{cases}$$

where each of z_i is independent and identically distributed according to a multinomial distribution of one drawn from K categories with probabilities $\tau_i, \dots, \tau_K$, the E-Step of EM iteration is given by

$$\hat{z}_{ik} = \frac{\hat{\tau}_k f_k(x_i \mid \hat{\boldsymbol{\theta}})}{\sum_{j=1}^{K} \hat{\tau}_j f_j(x_i \mid \hat{\boldsymbol{\theta}})} . \tag{1}$$

For the M-Step, estimates of the means and mixing probabilites have the following forms involving the data and $\hat{z}_{ik}$ from the E-Step:

$$\hat{\tau}_k = \frac{n_k}{n}, \qquad \hat{\mu}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ik} x_i}{n_k}, \qquad n_k = \sum_{i=1}^{n} \hat{z}_{ik} .$$

2.2 Testing the number of components with parametric bootstrap

Let $x_1, \dots, x_n$ be a random sample from a distribution with the probability density function $h(x)$. A test of

$H_0 : h(x) = h_0(x \mid \boldsymbol{\theta}_0)$ is a normal mixture with a common variance where dim $(\boldsymbol{\theta}_0) = k_0$
$H_1 : h(x) = h_1(x \mid \boldsymbol{\theta}_1)$ is a normal mixture with a common variance where dim $(\boldsymbol{\theta}_1) = k_1$
(where $k_0 < k_1$)

is considered and the likelihood ratio test statistic is defined as

$$-2 \log \lambda = -2 \sum_{i=1}^{n} \log \frac{h_0(x_i \mid \boldsymbol{\theta}_0)}{h_1(x_i \mid \boldsymbol{\theta}_1)} .$$

As mentioned in Section 1, the above statistic does not follow an asymptotic chi-squared distribution. In this section, we extend the bootstrap method by McLachlan (1987) to a mixture with more than two components for an application to the SULT1A1 data. In this method, a random sample of size n is repeatedly generated from a population with density $h_0(x \mid \hat{\boldsymbol{\theta}}_0)$, where $\hat{\boldsymbol{\theta}}_0$ is the MLE of $\boldsymbol{\theta}_0$ obtained from the original data. The test statistic is then calculated from each generated data set by fitting mixture models with k_0 and k_1 parameters and this constitutes the empirical distribution of a test statistic under null hypothesis. The bootstrap p-value can be obtained and is compared to size α for the decision.

2.3 Estimation of cancer risk with logistic regression

Let x_i denote the phenotype of a subject or an appropriate transformation of the phenotype, e.g., logarithm of it. Ostensibly, the rate of metabolic reactions modifies the carcinogenic risk of an exposure, and this suggests that x_i could be used as a covariate in case-control logistic regression. But x_i is a substrate specific measure and it may not correlate very well with the *in vivo* metabolism that is the direct source of risk. Another possible procedure is to assign each subject into a specific genotype based upon x_i. However, determination of the arbitrary cut-off point of the $x_{i'}$s is problematic and may lead to misclassification errors, especially when x_i does not resolve the genotype very well. One possible approach to estimating the cancer risk of putative genotypes is to replace an indicator vector for these genotypes in a standard logistic regression model with an expectation for this vector given the observed phenotype (Delongchamp, 1993). This expectation is closely related to the expectation that is calculated in EM algorithm for a mixture-of-genotypes model. Under this model, the likelihood of observed x_i given that the subject has genotype g is

$$\frac{1}{\sigma}\,\phi\left(\frac{x_i-\mu_g}{\sigma}\right).$$

Let $w_g(x_i)$ be a weight which is defined to be

$$w_g(x_i)=\frac{\hat{\tau}_g\phi\left(\dfrac{x_i-\hat{\mu}_g}{\hat{\sigma}}\right)}{\Sigma_g\hat{\tau}_g\phi\left(\dfrac{x_i-\hat{\mu}_g}{\hat{\sigma}}\right)}\,. \tag{2}$$

The weight function $w_g(x_i)$ can replace the indicator vector in logistic regression and this would perform better with less bias from misclassification by incorporating the possible contribution of each x_i to all the genotypes. The logistic model $\log\left(\dfrac{E[y_i]}{1-E[y_i]}\right)=\log\left(\dfrac{\pi}{1-\pi}\right)=\beta_0+\Sigma_g\beta_g w_g(x_i)$, where y_i is a responsible variable indicating colon cancer case for the i-th subject, can be employed to estimate the relative risk of colon cancer. ML estimators, $\hat{\beta}_0$ and the $\hat{\beta}_g$'s are the solution of the following esitmating equations and need to be obtained numerically.

$$\sum_{i=1}^{n} y_i=\sum_{i=1}^{n}\frac{e^{\hat{\beta}_0+\Sigma_g\hat{\beta}_g w_g(x_i)}}{1+e^{\hat{\beta}_0+\Sigma_g\hat{\beta}_g w_g(x_i)}}$$

$$\sum_{i=1}^{n} w_g(x_i)y_i=\sum_{i=1}^{n}\frac{w_g(x_i)e^{\hat{\beta}_0+\Sigma_g\hat{\beta}_g w_g(x_i)}}{1+e^{\hat{\beta}_0+\Sigma_g\hat{\beta}_g w_g(x_i)}}\,.$$

Also, the genotype frequency can be estimated by

$$\hat{p}_g=\frac{1}{n}\sum_{i=1}^{n} w_g(x_i)\,.$$

3 Results

3.1 Mixture model fitting

Figure 1 displays the histograms of SULT1A1 activities (phenotypes) in the colon cancer case and control study. To fit the mixture-of-genotypes model to the data, several issues need to be investigated such as whether the data are from a 3-component mixture or 4-component mixture of normal distributions or whether the means of the components shift in their magnitude when colon cancer cases are compared to controls. This is important because three genotypes can be explained with two alleles while at least three alleles are required to generate four or more genotypes. We calculate $-2\log\lambda$ for testing the null hypothesis of 3-component mixture versus the alternative of 4-component mixture. Further, we calculate $-2\log\lambda$ under the null hypothesis that the colon cancer case and control have the same means with different mixing proportions and the alternative that the colon cancer case and control have different means with different mixing proportions. The bootstrap samples under the null hypotheses in four possible null-alternative combinations are then generated. Figure 2 shows the em-

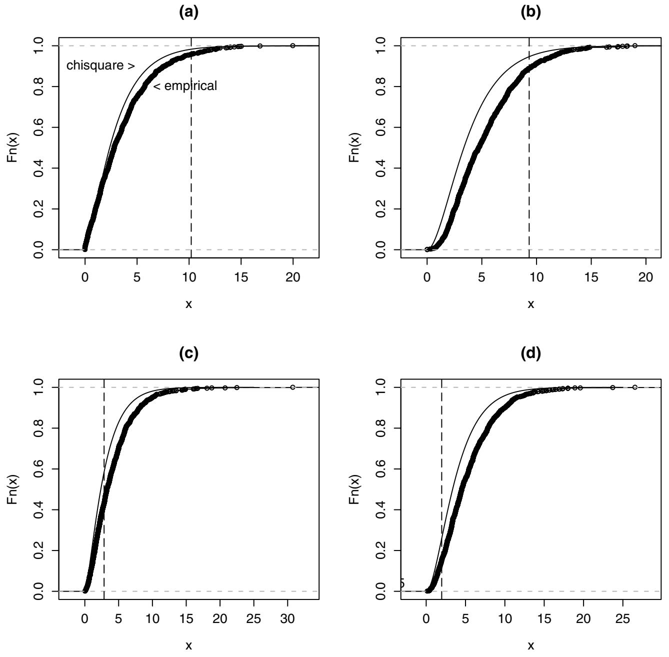

Figure 2 Empirical distributions of $-2\log\lambda$ for testing (a) H_0: 3-component mixture vs. H_1: 4-component mixture with the same means for the colon case and control (compared with the chi-squared distribution with d.f. = 3) (b) H_0: 3-component mixture vs. H_1: 4-component mixture with different means for the colon cancer case and control (compared with the chi-squared distribution with d.f. = 3) (c) H_0: the same means for the colon cancer case and control vs. H_1: different means for the colon cancer case and control with 3-component mixture (compared with the chi-squared distribution with d.f. = 3) (d) H_0: the same means for the colon cancer case and control vs. H_1: different means for the colon cancer case and control with 4-component mixture (compared with the chi-squared distribution with d.f. = 3).

Table 1 The value of $-2\log\lambda$ with bootstrap p-value – Colon cancer case and control. The p-value based on a chi-squared distribution is given in parentheses.

Hypothesis	Condition	$-2\log\lambda$	p-value
H_0: Same vs. H_1: Different	3-component mixture	2.850	0.570 (0.415 χ^2 d.f. 3)
means for case and control	4-component mixture	1.972	0.847 (0.741 χ^2 d.f. 4)
H_0: 3-component mixture	Same means for case and control	10.207	0.038 (0.017 χ^2 d.f. 3)
vs. H_1: 4-component mixture	Different means for case and control	9.328	0.111 (0.053 χ^2 d.f. 4)

pirical distributions of a thousand bootstrap replications compared with the chi-squared distribution in each case. Likelihood ratio test statistics obtained from the original data set with the p-value calculated from both the empirical distribution of bootstrap replications and the chi-squared distribution are given in Table 1. Although the decisions based on both distributions are the same at significance level 0.05, the empirical distribution tends to have larger p-values than the chi-squared distribution. The test for same means versus different means suggests the same means ($p = 0.847$). Under the assumption of the same means for case and control, the test prefers the 4-component mixture to the 3-component mixture ($p = 0.038$). Thus, the test results suggest the 4-component normal mixture with the same means for the colon cancer case and control.

Figure 3 presents a 4-component normal mixture fitting to the logarithm of the SULT1A1 phenotype for the colon cancer case and control. The parameter estimates using EM algorithm are given below.

	Case		Control	
$\hat{\mu}_g$	$\hat{\tau}_g$	$\hat{\sigma}^2$	$\hat{\tau}_g$	$\hat{\sigma}^2$
−3.57	0.05	0.26	0.06	0.18
−1.62	0.26		0.20	
−0.05	0.50		0.67	
1.07	0.19		0.07	

The distribution of SULT1A1 phenotypes in the prostate cancer case-control study is determined in a similar way. The underlying assumption for the model is that the data are from a normal mixture and the parameter estimation is performed using EM algorithm. Two-component normal mixture with the same means for prostate cancer case and control is chosen as a model for these data based on the parametric bootstrap likelihood ratio test where the results of which several possible scenarios are given in Table 2.

Tests for the same means versus different means for both one-component normal and two-component mixture do not reject the null hypothesis. Under the assumption of the same means for case and control, the test rejects the null of one-component normal and is in favor of the alternative of two-component mixture. Note that the test result for mixture with more than two components is not significant. The following table shows the parameter estimation of this model.

	Case		Control	
$\hat{\mu}_g$	$\hat{\tau}_g$	$\hat{\sigma}^2$	$\hat{\tau}_g$	$\hat{\sigma}^2$
−0.52	0.07	0.28	0.13	0.21
0.50	0.93		0.87	

3.2 The relative risk estimation with logistic model fitting

Under our model, we calculate $w_g(x_i)$ in Equation (2) of each subject and fit the logistic model using the $w_g(x_i)$'s as explanatory variables. For the colon cancer data, since $\Sigma_g w_g(x_i) = 1$, the model $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{g=1}^{4} \beta_g w_g$ is reduced down to

$$\log\left(\frac{\pi}{1-\pi}\right) = (\beta_0 + \beta_4) + (\beta_1 - \beta_4)\, w_1 + (\beta_2 - \beta_4)\, w_2 + (\beta_3 - \beta_4)\, w_3 = \beta_0' + \sum_{g+1}^{3} \beta_g' w_g$$

where β_g' represents the odds ratio of genotypes g and 4 ($g = 1, 2, 3$) in log scale. The p-values of the test for the null hypothesis that $\beta_g' = 0$ are 0.0431, 0.0535, 0.0018, for $g = 1, 2, 3$, respectively.

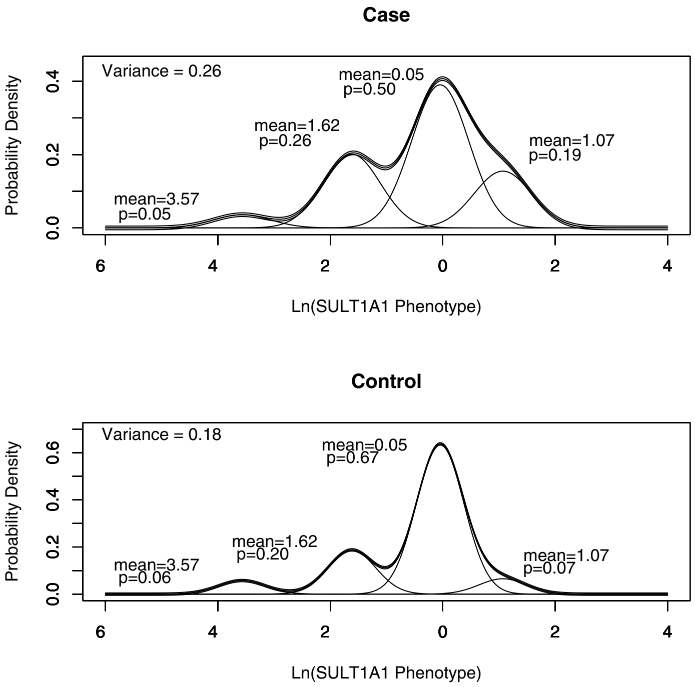

Figure 3 A 4-component mixture fitting to the logarithm of the SULT1A1 activities for the colon cancer case and control.

Similarly for the prostate cancer data, logistical model $\log\left(\frac{\pi}{1-\pi}\right) = \beta'_0 + \beta'_1 w_1$, where β'_1 is the odds ratio of genotype 1 and 2, is applied and the corresponding p-value is 0.003. The odds ratios and corresponding 95% confidence intervals obtained from the model are

		Estimate	95% CI
Colon cancer	Genotypes 1 and 4	0.129	(0.018, 0.928)
	Genotypes 2 and 4	0.252	(0.062, 1.021)
	Genotypes 3 and 4	0.112	(0.028, 0.443)
Prostate cancer	Genotypes 1 and 2	0.159	(0.047, 0.539)

Table 2 The value of $-2\log\lambda$ with bootstrap p-value – Prostate cancer case and control. The p-value based on a chi-squared distribution is given in parentheses.

Hypothesis	Condition	$-2\log\lambda$	p-value
H_0: Same vs. H_1: Different means for case and control	1-component normal	1.346	0.268 (0.246 χ^2 d.f. 1)
	2-component mixture	0.886	0.702 (0.642 χ^2 d.f. 2)
H_0: 1-component normal vs. H_1: 2-component mixture	Same means for case and control	9.472	0.029 (0.024 χ^2 d.f. 3)
	Different means for case and control	9.012	0.078 (0.061 χ^2 d.f. 4)
H_0: 2-component normal vs. H_1: 3-component mixture	Same means for case and control	3.722	0.338 (0.293 χ^2 d.f. 3)

Note that one can fix any genotype as a reference genotype whose odds will be modeled against the odds of other genotypes. In our application, we chose a genotype with the highest mean value of SULT1A1 activity, thus, genotype 4 for the colon cancer and genotype 2 for the prostate cancer. From the above result, we found non-homogenous odd ratios between genotypes, especially between genotypes 1 and 4, and genotypes 3 and 4 for the colon cancer and genotypes 1 and 2 for the prostate cancer case. The small values of the odds ratio (less than 1) suggest that the reference genotype has higher cancer risk than others. Based on our model, the cancer risk for an individual can easily be obtained by plugging in the weights ($w_g(x_i)$) of an individual into the fitted model.

We used $w_g(x_i)$ as a predictor variable which is estimated from mixture model. This may introduce a measurement error to the predictor variable and cause a bias to the estimated coefficient in the logistic regression. In the presence of additive measurement error, under-estimated slope with more variability is expected and proper adjustment may be necessary (Fuller, 1987; Carroll et al., 1995).

3.3 Simulation studies

3.3.1 A comparison of the parametric bootstrap likelihood ratio test and the standard χ^2 test

A simulation study is performed for evaluating the performance of the parametric bootstrap likelihood ratio test and compare it with the standard χ^2 test. Both the size and power evaluations are carried out with 5% significance level. In order to find out if the results obtained in Section 3.1 are reasonable, we generated simulation data sets which have similar features as the SULT1A1 data on colon cancer.

We considered two situations. First, a test for the null of 3-component normal mixture versus the alternative of 4-component normal mixture is conducted when the means are the same for the case and control. Second, a test for the null of the same means for the case and control versus the alternatvie of different means for the case and control is conducted when the data are from 4-component normal mixture.

In the size evaluation two hundred data sets of 46 cases and 121 controls are generated from the null distribution with the parameters estimated from the colon cancer data. Each simulated data set is then tested using parametric bootstrap likelihood ratio test with two hundred bootstrap replications. The summary of the simulation study of the size evaluation is given in Table 3. The chi-squared test is anticonservative for Test 1 while it controls size for Test 2. The parametric bootstrap likelihood ratio test controls the size as expected in both situations.

Table 3 Simulated size at 5% significance level based on 200 data sets with 200 bootstrap replications.

Test 1:	H_0	4-component normal mixture	same means for case & control
	H_1	4-component normal mixture	different means for case & control
Parameters	$\boldsymbol{\theta}_0$	$\boldsymbol{\mu}_0 = (-3.57, -1.62, -0.04, -1.07)$	
	Case	$\boldsymbol{p}_0 = (0.05, 0.26, 0.50, 0.19)$	$\sigma_0 = 0.51$
	Control	$\boldsymbol{p}_0 = (0.06, 0.20, 0.67, 0.07)$	$\sigma_0 = 0.42$
Size	Bootstrap	LRT	5.0%
	χ^2	test	11.5%
Test 2:	H_0	3-component normal mixture	same means for case & control
	H_1	4-component normal mixture	
Parameters	$\boldsymbol{\theta}_0$	$\boldsymbol{\mu}_0 = (-3.55, -1.54, 0.11)$	
	Case	$\boldsymbol{p}_0 = (0.04, 0.25, 0.71)$	$\sigma_0 = 0.75$
	Control	$\boldsymbol{p}_0 = (0.06, 0.21, 0.73)$	$\sigma_0 = 0.53$
Size	Bootstrap	LRT	2.5%
	χ^2	test	5.0%

Similarly, two hundred sets of 46 cases and 121 controls are generated from the alternative distribution with the parameters obtained from the colon cancer data for the power evaluation. With 4-component normal mixture, the result for testing the null of the same means for the case and control versus the alternative of different means for case and control (Test 1) is given in the first half of Table 4. In accordance to the result obtained in Section 3.1 that the original colon cancer data have the same means for the case and control, the parameters of the alternative distribution are not much distinctive. The power obtained from the bootstrap LRT is only 11%, while the power from the chi-squared test is 30%. The bootstrap LRT supports the results from Section 3.1 that the null hypothesis is not rejected. In order to obtain a higher power, we also generated 4-component normal mixture data with substantially bigger separation of the means. As expected, substantially larger power is observed from both the bootstrap LRT and chi-squared test.

When the means for case and control are the same, the power of the parametric bootstrap LRT and chi-squared test for testing the null of 3-component normal mixture versus the alternative of 4-component normal mixture (Test 2) is given in the second half of Table 4. The simulation data sets are generated from the alternative distribution (4-component mixture and different means for case and control) with parameter $\boldsymbol{\theta}_1$ obtained from the colon cancer data. The high power obtained in this simulation supports the conclusion from the analysis of the colon cancer data that the 4 components have enough separation.

3.3.2 Overall performance of the proposed approaches

We perform another simulation study to evaluate the overall performance of the proposed approaches. This simulation study is designed as follows. For the fixed genotype frequencies of cases for 4-compo-

Table 4 Simulated power at 5% significance level based on 200 data sets with 200 bootstrap replications.

Test 1:	H_0	4-component normal mixture	same means for case & control
	H_1	4-component normal mixture	different means for case & control
Parameters	$\boldsymbol{\theta}_1$		
	Case	$\boldsymbol{\mu}_1 = (-3.91, -1.67, -0.11, 0.83)$	
		$\boldsymbol{p}_1 = (0.04, 0.25, 0.43, 0.28)$	$\sigma_1 = 0.51$
	Control	$\boldsymbol{\mu}_1 = (-3.44, -1.57, -0.05, 1.15)$	
		$\boldsymbol{p}_1 = (0.07, 0.20, 0.67, 0.06)$	$\sigma_1 = 0.41$
Power	Bootstrap	LRT	11.0%
	χ^2	test	30.0% (d.f. 4)
Parameters	$\boldsymbol{\theta}'_1$		
	Case	$\boldsymbol{\mu}'_1 = (-4.57, -2.28, -0.12, 0.83)$	
		$\boldsymbol{p}'_1 = (0.04, 0.25, 0.43, 0.28)$	$\sigma'_1 = 0.51$
	Control	$\boldsymbol{\mu}'_1 = (-3.44, -2.57, 0.48, 1.95)$	
		$\boldsymbol{p}'_1 = (0.07, 0.20, 0.67, 0.06)$	$\sigma'_1 = 0.41$
Power	Bootstrap	LRT	64.0%
	χ^2	test	87.5%
Test 2:	H_0	3-component normal mixture	same means for case & control
	H_1	4-component normal mixture	
Parameters	$\boldsymbol{\theta}_1$	$\boldsymbol{\mu}_1 = (-3.57, -1.62, -0.04, 1.07)$	
	Case	$\boldsymbol{p}_1 = (0.05, 0.26, 0.50, 0.19)$	$\sigma_1 = 0.51$
	Control	$\boldsymbol{p}_1 = (0.06\ 0.20, 0.67, 0.07)$	$\sigma_1 = 0.42$
Power	Bootstrap	LRT	97.0%
	χ^2	test	98.5% (d.f. 3)

nent normal mixture which are the same as those in the colon cancer data set (0.05, 0.26, 0.5, 0.19), we set the true odds ratios between genotypes 2 and 4, and 3 and 4 as 2.0 and 0.22, respectively. By fixing the genotype frequency of the 4th genotype in controls as 0.07 which is the value from the colon cancer data set, we can obtain the genotype frequencies of other genotypes (0.045, 0.048, 0.837) and the odds ratio of genotypes 1 and 4 (0.41). One thousand data sets of 46 cases and 121 controls are generated from the 4-component mixture distribution with these obtained genotype frequencies, and the means and standard deviations from the colon cancer data set. We then applied the proposed approaches (mixture model and logistic regression fitting using w_g's as covariates) to each simulated data set and obtained the 95% confidence intervals of the odds ratio. The percentage of simulated data sets that converged to the 4-component mixture is 80.2% and among those converged to the 4-component mixture, 87.4%, 84.4% and 70.9% of the confidence intervals contained the true odds ratios for the genotype 1 and 4, 2 and 4, and 3 and 4, respectively.

4 Discussion

Many of the enzymes, which are involved in the metabolism of exogenous chemicals, have genetic variants that commonly occur in human populations. These variants can have quite different activities with respect to the metabolism of a specific substrate. Our hypothesis is that a subject's genotype for such metabolic enzymes modifies their carcinogen exposures and thereby alters their cancer risk. This hypothesis is being investigated through cancer case-control studies. When a genotype for each subject is known, estimating the relative cancer risk associated with the genotype is a straightforward application of logistic regression. Operationally this paradigm requires knowledge of the gene(s) which potentially modify risk. An alternative paradigm is to examine a phenotype distribution of a key metabolic enzyme for evidence of 'variants', especially looking for enzymes where the 'variant' distribution changes with case status. SULT1A1 activity illustrates the concept. Figure 1 displays a distribution of SULT1A1 activities with prominent modes that suggest the measured population is a mixture. Further, the relative proportions of the components appear to differ between the cases and the controls. This is at least superficially consistent with an underlying distribution of genotypes that modify cancer risk.

A purpose of this paper is to characterize the distribution of SULT1A1 activities as a mixture of unknown genotypes and to evaluate the disease risk associated with these putative variants. One would anticipate that the control subjects from the colon cancer study would exhibit a similar phenotype distribution to the control subjects in the prostate cancer study. This is not the case mostly because these studies employed different methods to assay SULT1A1 activity. In addition, the assay used in the colon cancer study is believed to measure the combined activity of SULT1A1 and another enzyme. Hence, the phenotype recorded in the colon cancer study differed substantially from the phenotype recorded in the prostate cancer study. While this example may not be the best, it does illustrate the somewhat arbitrary nature of such phenotype definitions and suggests the value of statistical methods that can relate the disease risk to underlying genotypes. In the analyses presented here, the phenotype distribution of SULT1A1 activities in a population is assumed to be a contribution from several genotypes. Recognition of this fundamental fact suggests that the observed distribution could be statistically modeled as the sum of probability distributions, one for each genotype weighted by its frequency in the population. In this paper, we focused on the colon cancer data because its phenotype distribution indicates several genotypes. The recorded SULT1A1 activities were fit by a mixture of four normal distributions, the putative genotypes. The model finds that cases differ significantly from controls in the mixing proportions of these four distributions (genotypes) but not in their means. Taken at face value, this indicates that the supposed genotypes affect the risk of colon cancer. Specifically, the genotype with a low frequency (next to the lowest) has greater impact on cancer risk than more frequent genotypes. The SULT1A1 activity (phenotype) corresponding to this genotype is the biggest and this suggests higher phenotype of this enzyme perhaps is important in studying the association between cancer risk and SULT1A1 activity. The result from prostate cancer data also supports this argument.

Table 5 Relationships among phenotype frequencies, genotype frequencies and allele frequencies (no linkage implies $P(B\,|\,A) = P(B\,|\,a)$).

Phenotype	Genotype	Frequency
1	AABB	π_{AB}^2 where $\pi_{AB} = P(A)\,P(B\,|\,A)$
	AABb	$2\pi_{AB}\pi_{Ab}$ where $\pi_{Ab} = P(A)[1 - P(B\,|\,A)]$
	AaBB	$2\pi_{AB}\pi_{aB}$ where $\pi_{aB} = [1 - P(A)]\,P(B\,|\,a)$
	AbBb	$2\pi_{AB}\pi_{aB} + 2\pi_{Ab}\pi_{aB}$ where $\pi_{ab} = [1 - P(A)][1 - P(B)]$
2	AAbb	π_{Ab}^2
	Aabb	$2\pi_{Ab}\pi_{ab}$
3	aaBB	π_{aB}^2
	aaBb	$2\pi_{aB}\pi_{ab}$
4	aabb	π_{ab}^2

If the four components correspond to distinct genotypes, then a simple two allele genetic model, one with high activity relative to the other, cannot explain these data. Of course, non-genetic explanations for an apparent mixture of four populations are possible, and they cannot be excluded by these methods. However, if there is a genetic basis for the observed distribution, then it should be consistent with four prominent modes.

Three alleles (A, B, C) imply six genotypes (AA, AB, AC, BB, BC, CC), but not four. However, with a sample size of 121 control subjects and 46 cases, rare genotypes might not be adequately represented for them to appear as distinct peaks. Another possibility is that two alleles are codominant and the third is recessive to the another two. An alternative model is that the SULT1A1 phenotype depends upon an additional gene, having two alleles. That is, four phenotypes can arise from two loci where each locus has dominant and recessive alleles. Table 5 lists the nine possible genotypes and their corresponding phenotype. We assume that the haplotypes, AB, Ab, aB and ab occur in the control population with relative frequencies, π_{AB}, π_{Ab}, π_{aB} and π_{ab}, respectively. The phenotype frequencies, τ_i, are simply the sum of the constituent genotype frequencies. For phenotype distributions associated with an enzyme, the two loci can reflect two single-nucleotide polymorphisms that alter the amino acid sequence. In these cases, haplotype frequencies are unlikely to reflect a random segregation of the alleles and the simplest genetic model can only estimate these frequencies. However, it is useful to rule out the random segregation hypothesis since it implies that the two loci are not linked. The probabilities of genotypes could be obtained from the mixing proportions using this alternative model.

According to our simulation study, the parametric bootstrap LRT tends to be more conservative than the chi-squared test. Although the chi-squared test appears to be more powerful than the parametric bootstrap LRT, the chi-squared test fails to control size in a certain case. Overall, the parametric bootstrap LRT controls the size quite well. Our simulation results support the conclusions made in the data analysis.

Here we attempt to classify subjects into ‘genotypes’ based on the phenotype. Alternatively, a phenotype measurement can be used directly as the covariate in logistic regression. However, the substrate being measured and the quantity defining a phenotype are somewhat arbitrary. In concept, they correlate with cancer rates through their ability to identify an underlying genotype, and it seems more appropriate to assess the data in that light, which led us to the modeling presented here.

The case-control study was originally designed to match controls to the sex, age and race of cases. Risk estimates for the genotypes should be adjusted for the resulting imbalance in these factors, especially race because genotype frequencies tend to vary substantially among the races. The effect of these factors on SULT1A1 activity may be incorporated by adjusting mean activity μ_g to $\mu_g\, e^{\boldsymbol{u}\boldsymbol{\beta}}$, where $\boldsymbol{u}$ is an indicator vector for these factors with corresponding parameter $\boldsymbol{\beta}$. The effect on genotype frequences may be done in a similar way, however, since this may lead to an over-parametrization, an alternative method that adjusts these factors in logistic regression may be desired. In the present study, these factors were not observed with many subjects, thus they could not be appropriately included in the modeling.

Acknowledgements The authors would like to thank the editor and two referees for their thoughful suggestions and comments which greatly improved our presentation.

References

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modeling of data on teaching styles. *Journal of the Royal Statistical Society. Series A* **144**, 419–461.

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31–38.

Böhning, D. (1998). *Computer-assisted analysis of mixtures and applications, Meta-analysis, disease mapping and others.* Chapman and Hall/CRC.

Boyle, C. R. and Elston, R. C. (1979). Multifactorial genetic models for quantitative traits in man. *Biometrics* **35**, 55–68.

Butler, M. A., Lang, N. P., Young, J. F., Caporaso, N. E., Vineis, P., Hayes, R. B., Teitel, C. H., Massengill, J. P., Lawsen, M. F., and Kadlubar, F. F. (1992). Determination of CYP1A2 and NAT2 phenotypes in human populations by analysis of caffeine urinary metabolites. *Pharmacogenetics* **2**, 116–127.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models.* Chapman and Hall, London.

Cox, D. B. and Hinkley, D. V. (1974). *Theoretical statistics.* Chapman and Hall, London.

Delongchamp, R. R. (1993). *Analysis of epidemiological data with covariate errors.* Ph.D. thesis, Oregon State University, Corvallis, OR

Dempster, A. P., Laird, N. M., and Rubin, D. R. (1977). Maximum likelihood from incomplete data. *Journal of the Royal Statistical Association. Series B* **39**, 629–646.

Efron, B. (1979). Bootstrap methods: Another look at the jacknife. *The Annals of Statistics* **7**, 1–26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* SIAM, Philadelphia.

Elston, R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21**, 523–542.

Frame, L. T., Ozawa, S., Nowell, S. A., Chou, H. C., Delongchamp, R. R., Doerge, D. R., Lang, N. P., and Kadlubar, F. F. (2000). A simple colorimetric assay for ehenotyping the major human thermostable phenol sulfotransferase (SULT1A1) using platelet cytosols. *Drug Metabolism and Disposition* **28**, 1063–1068.

Fuller, W. A. (1987). *Measurement Error Models.* John Wiley and Sons, New York.

Goffinet, B., Loisel, P., and Laurent, B. (1992). Testing in normal mixture models when the proportions are known. *Biometrika* **79**, 842–846.

Jackson, P. R., Tucker, G. T., and Woods, H. F. (1989). Testing for bimodality in frequency distributions of data suggesting polymorphisms of drug metabolism-histograms and probit plots. *British Journal of Clinical Pharmacology* **28**, 647–653.

Kaderlik, K. R. and Kadlubar, F. F. (1995). Metabolic polymorphisms and carcinogen-DNA adduct formation in human populations. *Pharmacogenetics* **5:S**, 108–117.

Lang, N. P., Butler, M. A., Massengill, J., Lawson, M., Stotts, R. C., Hauer-Jensen, M., and Kadlubar, F. F. (1994). Rapid metabolic phenotypes for acetyltransferase and cytochrome P4501A2 and putative exposure to foodborne heterocyclic amines increase the risk for colorectal cancer or polyps. *Cancer Epidemiology, Biomarkers & Prevention* **3**, 377–383.

Lo, Y., Mendell, N. R., and Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika* **83**, 767–778.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* **36**, 318–324.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models*. Marcel Dekker, New York.

Morton, N. E. and MacLean, C. J. (1974). Complex segregation analysis of quantitative traits. *American Journal of Human Genetics* **26**, 489–503.

Nowell, S., Sweeney, C., Hammons, G., Kadlubar, F. F., and Lang, N. P. (2002). CYP2A6 activity determined by caffeine phenotyping: association with colorectal cancer risk. *Cancer Epidemiology Biomarkers & Prevention* **11**, 377–383.

Soromenho, G. (1994). Comparing approaches for testing the number of components in a finite mixture model. *Computational Statistics* **9**, 65–78.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioural Research* **5**, 329–350.