INSULIN RESISTANCE: REGRESSION AND CLUSTERING

By

Sangho Yoon, Stanford University
Themistocles L. Assimes, Stanford University
Thomas Quertermous, Stanford University
Chin-Fu Hsiao, National Health Research Institutes
Lee-Ming Chuang, National Taiwan University
Chii-Min Hwu, National Yang-Ming University
Richard A. Olshen, Stanford University

Technical Report 248
August 2009

This research was supported in part by
National Institutes of Health grants R37EB02784 and 2HL05427-11
and by National Science Foundation grant CCR 0309701.

Division of Biostatistics
STANFORD UNIVERSITY
Sequoia Hall, 390 Serra Mall
Stanford, CA 94305-4065

http://statistics.stanford.edu
INSULIN RESISTANCE: REGRESSION AND CLUSTERING*

BY SANGHO YOON‡
THEMISTOCLES L. ASSIMES§
THOMAS QUERTERMOS¶
CHI-FU HSIAO**
LEE-MING CHUANG††
CHI-MIN HWU‡‡
RICHARD A. OLSHEN∥

Google‡, Stanford University‡§¶∥, National Health Research Institutes**, National Taiwan University††, National Yang-Ming University‡‡

This paper has two purposes. The first is to define insulin resistance (IR) precisely for a group of Chinese women. Our definition deliberately does not depend upon age and body mass index (BMI), although in previous studies that apply technologies we do not study here, BMI accounts for a large part of variability in IR. We accomplish our goal through careful application of Gauss mixture vector quantization (GMVQ), a technique for clustering that was developed for application to lossy data compression. Defining data come from measurements that play major roles in medical practice. They concern levels of lipids, and the results of an oral glucose tolerance test (OGTT). We apply GMVQ to residuals obtained from regression outcomes of the OGTT and lipids on functions of age and BMI that are inferred from the data. A bootstrap procedure developed for our family data leave us believing that two clusters are appropriate for our application, one cluster consisting of women who are IR and the other of women who seem not to be. The second purpose purpose of the paper is to examine the extent to which IR as we define it can be predicted from a combination of single nucleotide polymorphisms (SNPs) genotyped in particular “candidate genes” and also other environmental features. By design, age and BMI do not predict IR well. Further, age and BMI together with the “main effects” of SNPs in candidate genes are inadequate to describe IR much better than can

*Supported in part by NIH Merit Award R37EB02784, in part by NIH/NHLBI grant 2HL05427-11 to the University of Hawaiii with sub-allocation to Stanford University and in part by the National Science Foundation under NSF Grant CCR-0309701

AMS 2000 subject classifications: Primary 62H30, 62H30; secondary 62H30

Keywords and phrases: Insulin resistance, single nucleotide polymorphisms (SNPs), Gauss mixture vector quantization (GMVQ), feature selection, the bootstrap, CART, interactions, support vector machines (SVM), logistic regression, cross-validation, ROC curves

1
be predicted by chance alone. However, these features and their interactions enable somewhat accurate prediction in terms of area under an ROC curve. That in general IR is familial is patent, though its familiality in our data is complex and is analyzed by applying a conditional permutation statistic we develop. That the principal cause of IR is genetic is more subtle. We include a discussion of genome-wide association studies (GWAS) of complex (non-Mendelian) human disease and studies by SNPs in candidate genes, emphasizing difficulties with each approach.

1. Introduction. Insulin resistance results when glucose uptake by muscle and adipose tissue is sluggish, no matter levels of insulin being secreted by the pancreas. The underlying mechanism is hypothesized to be a pathway of glucose uptake by muscle and adipose tissue that is relatively compromised [41]. By way of compensation, beta cells in pancreatic islets hyper-secrete insulin in an attempt to maintain plasma glucose homeostasis. That effect is termed “hyperinsulinemia,” which is not a benign condition because of its negative effects on tissues that retain their insulin sensitivity such as the liver and kidneys. For example, insulin resistance/compensatory hyperinsulinemia clearly increases the production and excretion of triglycerides by the liver, resulting in elevated serum levels of triglycerides; and secondarily, it results in decreased levels of high density lipoprotein cholesterol (HDL) as well as increased presence of smaller and more than normally dense low density lipoprotein (LDL) particles. Insulin resistance/compensatory hyperinsulinemia also appears to contribute to the complex pathogenesis of the most common form of elevated blood pressure, namely essential hypertension [40], by promoting water resorption in the kidney and/or increasing activity of the sympathetic nervous system. Through these and other “downstream” adverse metabolic consequences whose description is beyond the scope of this brief introduction, insulin resistance/compensatory hyperinsulinemia markedly increases the risk of developing type 2 diabetes and the various manifestations of atherosclerotic cardiovascular disease, including myocardial infarction, ischemic stroke, and peripheral arterial disease. Current conventional wisdom has it that an individual’s degree of insulin resistance can be estimated biologically most accurately by one of two procedures. The euglycemic clamp is one; the insulin suppression test the other. We utilize results of the latter, “steady state plasma glucose,” or SSPG, in understanding inference for some subjects. Continuous measurements that result from these two procedures are highly correlated (.92) [24]. Studies by which insulin resistance is estimated using such “gold standard” measures confirm a very wide range of insulin sensitivity in healthy, non-diabetic,
non-hypertensive individuals. However, estimating insulin resistance using either of these technologies is invasive and laborious, has not made its way to the clinic thus far and is unlikely to in the foreseeable future. Multiple approximations have been developed. They correlate modestly (.5 to .7, [24]) with “gold standard” measures. Indeed, the approximations involve systemic background levels of glucose and insulin that are calibrated with the gold standards. SSPG is one measure of the body’s ability to respond to a fixed glucose challenge with a fixed level of insulin made available, whereas the approximations we study for all subjects entail at least two important background nuisance parameters for each person (insulin and glucose). Since IR is inferred largely from higher-than-expected values of measured quantities such as glucose and insulin following a glucose challenge, and since “errors of measurement” are well known to attenuate “real” effects, one might expect that IR as defined by surrogate measures underestimates actual incidence of IR by individual. This is borne out by data available to us. Cited attenuation is well known in the social sciences and follows easily from “wide sense” conditional models [10][20].

Surrogates involve easily obtainable fasting measures of serum insulin, glucose, triglyceride, and HDL levels, and measures of glucose and insulin over the two hour window of time after a glucose challenge. Measures of glucose and insulin at fasting baseline and after a glucose challenge comprise an “oral glucose tolerance test,” or OGTT. Results of an OGTT and measures of lipids have been taken to define “insulin resistance,” albeit subjectively. Previous “definitions” of insulin resistance involve high triglyceride to HDL ratio, a homeostatic model assessment of insulin resistance (termed HOMA IR), and high values of insulin area under the curve (AUC) for a period following a glucose challenge. Two powerful determinants of IR are adiposity as estimated by body-mass index (BMI) or waist circumference, and physical fitness as quantified by maximal oxygen uptake with exercise (VO$_2$ max). These two factors are in principle modifiable within individuals; obviously, they tend to be negatively correlated. By conventional analyses they may explain up to 50% of gold standard measures of insulin sensitivity in linear models [27][5][43].

Analyses of this paper are of data derived from 557 Chinese women from our SAPPHIRe study of hypertension, for which IR was taken as an intermediate, that is to say predisposing, phenotype. Chinese women are ideal subjects for studying genes that predispose to hypertension and their synergistic effects with other features because hypertension is somewhat prevalent among Chinese; typically they are not IR because they are obese; and both environmental and genetic factors are clearly more homogeneous among
them than among any general population.

Our goal has been to study the impact of genes and other features as they bear upon IR above and beyond what might be ascribed to age and adiposity, which were mentioned to be predictive. Age per se is believed to have little effect on IR, and that is validated here; but not so BMI. We took great pains to remove their effects. Seventeen functions of age and BMI are reduced to two principal components, which depend upon the five functions of age and BMI that devolve first from Taylor expansions of their values. We use a bootstrap technique derived specifically to handle SAPPHIRe family structures to deduce how many principal components and which specific functions that determine them are candidates for inclusion in further analyses. The residuals from a regression of other medical measures on chosen functions of age and BMI are clustered by what in information theory is termed Gauss mixture vector quantization (GMVQ) [2]. Bootstrapping leads to our choice of two clusters, one smaller group clearly IR and the other not. The clusters enable a precise definition of IR, not given heretofore for our subjects. We mention without drilling deeply to understand why that our definition of IR, based as it is on surrogate markers, does not always agree with SSPG on the subset of subjects for whom the latter was available.

What SAPPHIRe is and how it fits as a network of the (U.S. National Heart, Lung and Blood Institute’s) Family Blood Pressure Program (FBPP) are given in the next section. We explain how SAPPHIRe data were gathered so as to preclude familiality of hypertension or its close relative IR. We state why, and give a precise permutation-based test of familiality given our cluster-based definition. Unlike the usual situation, that hypertension and IR could not be familial in SAPPHIRe does not a priori preclude their etiologies being purely genetic, purely environmental, or a result of genes by environment interactions.

We had 293 single nucleotide polymorphisms (SNPs) genotyped in a total of 57 “candidate genes,” the respective proteins for which they code bearing upon hypertension. In particular, many such genes were chosen by resident SAPPHIRe experts because their proteins influence blood pressure control or glucose homeostasis. For example, APOAV regulates triglyceride levels, which are known to be differentially expressed in insulin resistance. CD36 is a multifunctional receptor, which plays a part in mediating intracellular signalling as well as in taking up biologically active substances such as long-chain fatty acids. Primers were designed to sequence the promoter region, the 5' and 3' untranslated regions (UTRs), the exons, and the intron-exon boundaries of each candidate gene. A discovery set comprised of 24 SAPPHIRe individuals’ DNA was sequenced to identify SNPs and
insertion-deletion mutations. These individuals were chosen to be hypertensive. After assembling the sequence contigs using the program Consed [22], the SNPs were tagged [30] and called manually in each of the 24 individuals. The SNPs identified in this manner were cross-checked against the public dbsNP database [44] and entered into a hand-curated SNP report. In general, SNPs not in high linkage disequilibrium (LD) with each other, that have greater than approximately a 10% allele frequency (see Table 14), and that were deemed likely to change protein function were chosen for genotyping. Genotyping was performed using the ABI Taqman 5' nuclease allelic discrimination system with either custom made or commercially available primers and probes. The accuracy of the genotypes was tested by comparing them against the discovery set sequences and against a 15% repeated set of DNA.

We applied five algorithms (support vector machine (SVM) [9][48], L1 (Lasso) regularization path algorithm for generalized linear models [38], logistic regression with L2 penalty [39], FlexTree [29], and random forests [6]) for predicting IR, by which we mean cluster membership, from genes, some aspects of the environment, and their interactions. These algorithms were chosen because they are popular currently in GWAS or have proven useful in previous studies of genetics by association (type 2 learning). While we argue informally that prediction of cluster membership on the basis of SNPs in candidate genes and other features (that did not figure in the clustering) is better than could be expected by chance, in fact our best algorithm, a support vector machine, is not sufficiently accurate for routine clinical application. Instead, to the extent that genotype, other features, and their synergistic effects predict IR, it may be better to have data from a GWAS than from candidate genes. This view is despite certain knowledge that any current GWAS necessarily entails beginning with not fewer than 500,000 features, most of them irrelevant, not to speak of the problem of describing “phenotype” (the dependent variable in prediction) accurately. Further, previous analyses of GWAS (see [49][33][13][17] for example) have focused primarily upon individual effects, no matter how minor their impact upon phenotype. This approach seems a vestige of thinking about Mendelian mechanisms of inheritance that do not port to this context of complex disease. Indeed, the argument for presenting our analyses is to demonstrate how statistics, broadly construed, can be brought to bear upon understanding prediction of complex human disease, and to show that despite the great care we took to define phenotype and to predict it as well as could be, information in “candidate genes” was simply not good enough in a family-based study like SAPPHIRe. Knowing proteins (in particular, genes that code for them) that
figure in clinical presentation of a phenotype such as insulin resistance is not the same as knowing also what controls the expression or other aspects of those genes/proteins. Such control may depend on genes far removed on the genome from those that code for the particular proteins. At present writing, at least so far as IR is concerned these other genes tend to be unknown. Our approach in some places is not adaptive; that is, it is unsupervised. However, this choice does not explain predictive performance.

2. SAPPHIRE data set. The FBPP is a large, multi-center study by genetics and other features of high blood pressure and related conditions in multiple racial groups. The program consists of four networks: GenNet, GENOA, HyperGEN, and SAPPHIRE [16]. Each has been funded by the U.S. National Heart, Lung, and Blood Institute (NHLBI) since 1995. FBPP focus has been to identify genes that contribute to essential hypertension or related phenotypes through genetic studies by linkage and association. For SAPPHIRE, IR was chosen as intermediate; that is to say predisposing, phenotype.

SAPPHIRE recruited Chinese and Japanese hypertensive patients in two phases. The first, from which data here were derived, were from four hospitals in Taiwan, one in Hawaii, and one in the San Francisco Bay Area. In this first phase, a total of 1460 Chinese siblings (602 males and 858 females) from 557 families were enrolled. Through an interview, a physical examination, a blood draw and an OGTT, 11 measurements (listed in Table 1) relevant to insulin resistance were obtained on all subjects. DNA was purified from whole blood and was used to genotype 293 SNPs in candidate genes as cited. From the original sample we excluded 517 people who were missing at least one crucial measurement. After exclusion there were 943 sibs, 386 males and the 557 females, for whom results are presented here. SSPG was also obtained for a subset of 202 female participants, but this subset was deemed too small to study exclusively. The final sample used to study insulin resistance in this work consists of 557 Chinese women who do not have any missing medical measurements. We do not dwell here on a matter that bears upon statistical inference. Thus, each sibship studied contained one proband for hypertension, that is, a subject who presented as hypertensive. Sibships involved hypertensive and other (typically hypotensive) sibs of the proband. SAPPHIRE data analyzed here were not gathered from a population study. Of course, this concern for conditional rather than unconditional attained significance and related applies to many epidemiological studies.

The approach to identify novel genetic determinants of insulin resistance is summarized here. First, we regress out the effect of all known determinants
of insulin resistance from measures commonly used to derive surrogates of insulin resistance including Triglyc, TCHL, HDL, OGTTG0, OGTTG1, OGTTG2, OGTTI0, OGTTI1, OGTTI2. Unfortunately, maximum exercise VO$_2$ was not measured in the SAPPHIRe study. Previous unpublished work has shown that 24 hour physical activity explains little if any of the variability of the insulin suppression test measure, which is available for only a subset of participants of the SAPPHIRe study. In any case, we regress out only the effects of BMI and age from these nine variables.

2.1. Scatter plots of variables. A scatterplot matrix of the measurements in Table 1 is shown in Figure 1 and demonstrates strong linear relationships between pairs among \{OGTTG0, OGTTG1, OGTTG2\}, and pairs among \{OGTTI0, OGTTI1, OGTTI2\}. Such relationships are expected on biological grounds.

3. Pre-processing data.

3.1. Residuals from linear regression. Age and BMI are known to predispose to IR. In this subsection we have two goals. The first is to define IR precisely using ideas from data compression combined with sample reuse methods. The definition for our group of Chinese women is “above and beyond” what could be inferred from age and BMI. Chosen functions of them were “subtracted” from numbers derived from an OGTT and from lipids, and the definition proceeded from what remained. Thus, we describe the IR phenotype as independent, functionally and statistically, to the extent we can. Having defined IR in this “age and BMI-free way,” we sought combinations of SNPs in candidate genes for hypertension, and age and BMI, that
enabled prediction of IR. Table 2 is a summary of 17 variables computed from age and BMI by which we attempted to capture their joint variability.

We sought to explain a large fraction of the variability in the joint distribution of age and BMI by principal components in the hope that a few principal components explain a large fraction of joint variability, and further that each principal component can be represented well by only a few of the 17 features listed in Table 2. Thus, principal components were computed on what amount to correlations of these 17 features. For each subject and each listed feature, that feature was divided by the standard deviation across subjects for that feature. Subsequent analyses were based on these standardized features. Subjects were bootstrapped respecting family structures, and “eigenvalue ratios” (what is sometimes called a “Scree plot”) was then fit for each of the 1,000 bootstrap samples. The simple results of the eigenvalue ratios/Scree plots and their constancy across bootstrap samples is described in Figure 2. As is seen from that figure, about 99% of variability is explained by the two largest eigenvalues in all 1,000 bootstrap samples.

Figure 3 shows how simply the first two of the cited eigenvalues can be described. Thus, $\text{AGE}^2$, $\text{BMI}^2$, and $\text{AGE} \times \text{BMI}$ were chosen as the tree features that summarize the joint variability in the joint distribution of age and

---

**Fig 1. A scatterplot matrix of eleven medical measurements in Table 1**

---

---

---

---
BMI. These three were independent variables in a simple linear regression of each of the nine medical measurements that quantify IR. The residuals from these nine regressions were then the ingredients to our clustering by which we chose subjects who seem to be IR and those who seem not to be.

We employ here a version of the patently “non-adaptive” approach PCA regression, and critics may ask what features of age and BMI we would have chosen with an adaptive regression. By this, we mean that the response variables were not considered in feature selection, though they might have been. To answer this ourselves, we used LARS [14] as an adaptive feature selector. For each of the nine medical features, LARS was run; and we examined a list of functions of age and BMI in the order selected by LARS. For six of the nine, the first three features selected were $\text{AGE}^2$, $\text{BMI}^2$, and $\text{AGE} \times \text{BMI}$. For eight of the nine, LARS chose these three functions of age and BMI among the first five features it chose. We conclude that the three functions of age and BMI we chose are reasonably supported by the data, no matter adaptive or non-adaptive regression.

3.2. Bootstrap. We mentioned bootstrapping that respects family structures and report here on methodology we developed for that purposes. Note that we cannot simply sample individually because data gathered from siblings in the same family are patently dependent. We sample by families so as to respect family structures. Table 3 shows sibship sizes in our SAPPHIRE

### Table 2

<table>
<thead>
<tr>
<th>index</th>
<th>functions of age and bmi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\text{AGE}$</td>
</tr>
<tr>
<td>2</td>
<td>$\text{BMI}$</td>
</tr>
<tr>
<td>3</td>
<td>$\text{AGE}^2$</td>
</tr>
<tr>
<td>4</td>
<td>$\text{BMI}^2$</td>
</tr>
<tr>
<td>5</td>
<td>$\text{AGE} \times \text{BMI}$</td>
</tr>
<tr>
<td>6</td>
<td>$\sqrt{\text{AGE}}$</td>
</tr>
<tr>
<td>7</td>
<td>$\sqrt{\text{BMI}}$</td>
</tr>
<tr>
<td>8</td>
<td>$\log(\text{AGE})$</td>
</tr>
<tr>
<td>9</td>
<td>$\log(\text{BMI})$</td>
</tr>
<tr>
<td>10</td>
<td>$\sin(\text{AGE})$</td>
</tr>
<tr>
<td>11</td>
<td>$\sin(\text{BMI})$</td>
</tr>
<tr>
<td>12</td>
<td>$\cos(\text{AGE})$</td>
</tr>
<tr>
<td>13</td>
<td>$\cos(\text{BMI})$</td>
</tr>
<tr>
<td>14</td>
<td>$\sin(\text{AGE}/2)$</td>
</tr>
<tr>
<td>15</td>
<td>$\sin(\text{BMI}/2)$</td>
</tr>
<tr>
<td>16</td>
<td>$\cos(\text{AGE}/2)$</td>
</tr>
<tr>
<td>17</td>
<td>$\cos(\text{BMI}/2)$</td>
</tr>
</tbody>
</table>
Chinese women. We denote the number of families with sibships (female) of size $j$ by $F_j$, and we denote the total number of individuals who belong to families of sibship size $j$ by $n_j$. Necessarily,

$$n_j = j \times F_j.$$  

The total number of individuals was

$$n = \sum_{j=1}^{6} n_j = \sum_{j=1}^{6} j F_j,$$  

(2)
and number of families was
\[ F = \sum_{j=1}^{6} F_j. \]  
(3)

From Table 3, there were 287 (\(= F\)) families and 557 (\(= n\)) individuals.

Bootstrapping was performed so that these three conditions were met:

- The expected number of families in a bootstrap sample was \( F \).
- The expected number of people in a bootstrap sample was \( n \).
- The expected fraction of families of size \( j \) in a bootstrap sample was the empirical fraction of families of size \( j \) in our SAPPHiRe data set.

To this end, bootstrap sampling was performed in two stages. Let \( J \) represent the sibship size of a family, \( J \) always an integer between 1 and 6. In keeping with the “bootstrap principle” that the bootstrap sample bears the same relationship to “the empirical distribution” that it bears to “nature,” we assume that

\[ \Pr(J = j) = \frac{F_j}{F}. \]  
(4)
Following the distribution in (4), we randomly generate $F$ sample values of $J_i$: $J_1, J_2, \ldots, J_{F-1}, J_F$. Each sample value $J_i$ can be viewed as a sampling of a family of sibship size $J_i$. Thus, we can define another variable to represent the fraction of families of sibship size $j$ in our sampled set:

\[
\hat{F}_j = \frac{1}{F} \sum_{i=1}^{F} I(J_i = j),
\]

where $I$ is the obvious indicator function. Once we have $\hat{F}_j$ for each $j$, we next sample $F \times \hat{F}_j$ families (with replacement) among families of sibship size $j$. It is obvious that this two-step sampling process satisfies the three conditions mentioned. Table 4 summarizes notations for population and bootstrapped sample.

The sampling approach in this section bears superficial resemblance to the famous Ewens sampling formula (see [15] and [32]). For example, in the notation of [28], think of $a_i$ as the number of sibships with $i$ members, what here we have called $F_i$; and see the summation following (1) in Section 2 of [32]. However, if we take the point of view of [28] and think of the Ewens sampling formula for the sequence of family sizes as arising from an urn scheme, then one sees that the Ewens formula corresponds more to what David Freedman called Bernard Friedman’s urn [18] than it does to the iid sampling scheme of our bootstrap. In particular, the population size is fixed in references [15][32], and [28]; but it is random here. For us, only the expected size of the population is fixed.

4. Clustering data. Vector quantization (VQ) design [21] amounts to clustering data. In VQ an input vector is represented by one of a predefined set of patterns (cluster centers = codewords) on the basis of which pattern is closest to the given input vector. VQ has been used successfully in pattern recognition, including speech and image processing [8][50][2]. VQ design can
Table 4

<table>
<thead>
<tr>
<th>Description</th>
<th>Population</th>
<th>Bootstrap sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of people of families of size (j)</td>
<td>(n_j)</td>
<td>(\hat{n}_j)</td>
</tr>
<tr>
<td>Fraction of families of size (j)</td>
<td>(F_j)</td>
<td>(\hat{F}_j)</td>
</tr>
<tr>
<td>Total number of people</td>
<td>(n)</td>
<td>(\hat{n})</td>
</tr>
<tr>
<td>Total number of families</td>
<td>(F)</td>
<td>(\hat{F})</td>
</tr>
<tr>
<td>Fraction of people of families of size (j)</td>
<td>(n_i/n)</td>
<td>(\hat{n}_i/\hat{n})</td>
</tr>
</tbody>
</table>

be also viewed as fitting a model when partition cells are represented by their conditional probability density functions, and prior probabilities are weights. VQ of dimension \(p\) (i.e., the number of features is \(p\)) and size \(K\) (the number of clusters is \(K\)) can be described by the mappings and sets: an encoder \(\alpha\), a decoder \(\beta\), and a length function \(l\). An encoder \(\alpha\) is a mapping of an input vector \(x\) in \(p\)-dimensional Euclidean space, \(\mathbb{R}^p\), into an index \(i \in \mathcal{I} = \{1, 2, ..., K\}\). The encoder is described by a partition \(S = \{S_i : i = 1, 2, ..., K\}\) such that \(S_i = \{x : \alpha(x) = i\}\). A decoder \(\beta\) converts the index into a source reproduction \(\hat{x}\), and \(\beta\) is associated with a reproduction codebook \(C = \{\beta(i) : i \in \mathcal{I}\}\). Finally, a length function \(l\) measures the complexity or cost of an index \(i\), and it is “admissible” if \(\sum_{i \in \mathcal{I}} e^{-l(i)} \leq 1\). Both \(l\) and the requirement of admissibility will be seen to be closely related to the “prior probability” of the cluster indexed by \(i\). For a fixed-rate quantizer, \(l(i)\) is fixed at \(\lceil \ln K \rceil\) (the integer part of \(\ln K\)) for all \(i\). Otherwise, a quantizer is said to be variable-rate. Eq. (6) summarizes VQ.

\[
X \overset{\alpha}{\rightarrow} \alpha(X) = i \overset{\beta}{\rightarrow} \hat{X} = \beta(\alpha(X))
\]

Here, for purposes of defining insulin resistance on the basis of certain residuals, we are interested in GMVQ, where we fit Gauss mixture models (GMM) to data in VQ design using the Lloyd algorithm with a suitable distortion measure [26][2]. The EM algorithm [12] is the most popular approach to fitting a GMM to data, but the Lloyd algorithm is one alternative. The main difference between the Lloyd and the EM algorithms is that in most implementations EM makes soft decisions whereas the Lloyd algorithm makes hard decisions. The EM fits a GMM to each observed vector, whereas the Lloyd fits a single component of a GMM to each observed vector. This “hard” assignment of components to observed data is based on the information theoretic property of a Gaussian being a “worst case” for designing robust compression/source coding systems [2][23].

In GMVQ, each cluster is represented by its prior probability \(w_i\) (\(w_i \geq 0\) and \(\sum_{i=1}^{K} w_i = 1\)) and a cluster conditional pdf \(g_i(x)\), a multivariate Gauss-
Fig 4. Top: $\hat{\rho}_{\text{GMVQ}}(q)$ vs. number of clusters of 1,000 bootstrapped samples. Bottom: $\Delta$ (decrease in $\hat{\rho}_{\text{GMVQ}}(q)$) vs. number of clusters of 1,000 bootstrapped samples.

Gaussian:

$$g_i(x) = g(x|\alpha(x) = i)$$

$$= \frac{1}{(2\pi)^p/2|\Sigma_i|^{1/2}} \exp \left( -\frac{(x - \mu_i)^T\Sigma_i^{-1}(x - \mu_i)}{2} \right)$$

where $g$ is a fitted GMM, and $\mu_i$ and $\Sigma_i$ are the mean vector and covariance matrix of cluster $i$, respectively; we assume $\Sigma_i$ to be non-singular.

In GMVQ, we try to minimize the mismatch between the true pdf $f$ and the fitted model $g$ by iterating the Lloyd optimality conditions. The encoding rule (or cluster assignment rule) is to find a component $g_i$ that minimizes the distortion $d_f(x, i) = \ln(f(x)/g_i(x)) - \ln(w_i)$. Since $\ln(f(x))$ is common to all $g_i(x)$, the encoding rule becomes:

$$\alpha(x) = \arg\min_i [-\ln(g_i(x)) - \ln(w_i)]$$

$$= \arg\min_i \left[ \frac{1}{2}(x - \mu_i)^T\Sigma_i^{-1}(x - \mu_i) + \frac{1}{2}\ln((2\pi)^p|\Sigma_i|) - \ln(w_i) \right].$$
When the true pdf $f$ is a GMM, minimizing the distortion $d_f(x,i)$ is equivalent to a maximum a posteriori selection (MAP) of a Gauss model from a GMM (a collection of Gauss models $g_i$ with a probability mass function $w_i$) [2]. The MAP selection of $g_i$ is

$$\alpha(x) = \text{argmax}_i P(i|x)$$

$$= \text{argmax}_i g_i(x)w_i$$

$$= \text{argmin}_i [-\ln(g_i(x)) - \ln(w_i)].$$

In GMVQ, we denote $\beta(i)$ (=centroid of $i^{th}$ cluster) by $N(\mu_i, \Sigma_i)$, equivalently $(\mu_i, \Sigma_i)$.

The distortion measure between $x$ and $\beta(i)$ can be expressed:

$$\rho_{\text{GMVQ}}(x,i) = d_{\text{GMVQ}}(x,\beta(i)) + l(i)$$

where $d_{\text{GMVQ}}(x,\beta(i)) = \frac{1}{2}(x - \mu_i)^t\Sigma_i^{-1}(x - \mu_i) + \frac{1}{2}\ln((2\pi)^p|\Sigma_i|)$, and $l(i) = -\ln(w_i)$. Finally the average distortion in GMVQ becomes:

$$\rho_{\text{GMVQ}}(f,q) = E_f(d_{\text{GMVQ}}(x,\beta(\alpha(X)))) + E_f(l(\alpha(X))).$$

When the underlying pdf $f$ is unknown, which is often the case in practice, the expectations in (10) become sample averages if an empirical distribution is used in the expectation:

$$\hat{\rho}_{\text{GMVQ}}(q) = \frac{1}{N} \sum_{i=1}^{K} \sum_{\alpha(x_n) = i} [d_{\text{GMVQ}}(x_n,\beta(i)) + l(i)],$$

where $N$ is the number of samples.

In GMVQ, the optimal $\beta(i)$ for a given encoder $\alpha(x)$ is defined by $\mu_i = E[X|\alpha(X) = i]$ and by $\Sigma_i = E[(X - \mu_i)(X - \mu_i)^t|\alpha(X) = i]$. The optimal length function in GMVQ is $l(i) = -\ln(w_i)$, where $w_i = Pr(\alpha(X) = i)$. See [23] [2] for more details. In practice, the conditional expectations become conditional sample averages when we run the Lloyd clustering algorithm on a finite set of samples. To avoid singular covariance matrices, regularization as in [19] can be used.

GMVQ was used to cluster people. It has been shown to perform well in many areas [1][2][51][37][50]. When GMVQ was applied to cluster people as was mentioned, the clustering was based on residuals from linear regression models described earlier. Insulin resistance was defined by clustering people, and the clustering model was validated internally. Therefore, we are trying to
solve an unsupervised learning problem: we not only need to cluster people, but also want to estimate the true number of clusters.

We tried first to estimate number of clusters by observing the GMVQ distortion in (11) as the number of clusters varied. Bootstrapping was performed 1,000 times. Figure 4 is a summary of results. The top plot in Figure 4 shows the sampling distribution of \( \hat{\rho}_{\text{GMVQ}}(q) \) as the number of clusters increased from one to eleven, and the bottom plot in Figure 4 represents decremental change in \( \hat{\rho}_{\text{GMVQ}}(q) \) as the number of clusters increased from one to eleven. From the figure, we see that the deepest decrease in \( \hat{\rho}_{\text{GMVQ}}(q) \) occurred when we increased the number of clusters from one to two. The GAP statistic [47] and the silhouette method [34] were also used to estimate the number of clusters; all confirmed that there are two clusters. We conclude that there are two clusters.

After clustering people into two clusters, means and standards deviation of the eleven medical measurements in Table 1 were measured for each cluster. See Table 5.

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|l|}
\hline
 & cluster1 (177 samples) & & cluster2 (380 samples) \\
 & mean & std & mean & std \\
\hline
Age & 50.60 & 8.57 & 48.94 & 8.50 \\
BMI & 25.74 & 3.51 & 24.14 & 3.45 \\
Triglycerides & 155.66 & 97.61 & 92.28 & 38.01 \\
Total cholestrol & 198.82 & 45.38 & 185.33 & 36.85 \\
HDL & 43.47 & 10.36 & 49.81 & 12.64 \\
OGTT glucose t=0 & 97.16 & 15.27 & 87.18 & 9.38 \\
OGTT glucose t=60 & 204.17 & 44.29 & 157.69 & 39.04 \\
OGTT glucose t=120 & 167.02 & 55.23 & 131.59 & 32.45 \\
OGTT insulin t=0 & 10.79 & 6.29 & 5.86 & 2.93 \\
OGTT insulin t=60 & 121.35 & 81.75 & 53.18 & 24.88 \\
OGTT insulin t=120 & 110.54 & 78.37 & 49.30 & 30.26 \\
\hline
\end{tabular}
\caption{Cluster statistics}
\end{table}

4.1. Model validation. This section is a summary of differences in the two clusters. While clustering and assignment of learning sample vectors to clusters were done on the basis of residuals of nine features after our attempt to remove the effects of age and BMI, we nonetheless can and do summarize clusters on the basis of the original 11 measurements. Were we successful in defining clusters that exist in our population of Chinese women, age and BMI should not differ by cluster. The other nine features “should.” Readers note that age hardly differs by cluster whereas BMI seems to. Further, the six measures of glucose and insulin are quite different between clusters, whereas
measures of lipids differ much less.

To be precise here, we report on what amount to approximate tests of significance between clusters performed thus.

- We bootstrapped data as cited; clustered the bootstrapped data; and performed a permutation pooled t-test, feature by feature.
- We also bootstrapped, analogous to what was done for the test of the previous bullet, except the test was the Mann-Whitney procedure.

We realize that permutation two-sample t is computed under assumptions that are obviously not true from cursory look at the Table 5. For the test to be valid by ordinary strict criteria of invariance, it should be that among permutations of bootstrap samples, only means by cluster differ, while variances are “the same.” One can show that were the numbers of individuals in each of the two clusters equal, this difficulty would not matter. Neither is so, and to that extent permutation t should be taken with a grain of salt as it applies to differences in means by cluster; but that comparison somehow might be viewed as unconditional with respect to an unspecified mechanism that generated equal intrinsic variability per feature per (independent) subject. At the same time, given that the marginal distributions of features in one cluster are in no cases merely location shifts of respective distributions in the other, it is not clear precisely what the Mann-Whitney null hypothesis should be. These being said, we think of each test as merely a test of the nonparametric hypothesis of equality of sampling distributions by subject, separately for each feature versus the alternative of inequality.

For t-testing 100 bootstrap samples of families were generated. For each bootstrap sample we performed 10,000 permutations of cluster membership, always keeping the total numbers of individuals per cluster constant. For each permuted data set, a t-statistic with pooled estimate of variance was performed (though Behrens-Fisher-t produced almost the same results). Then the p-value for the null hypothesis was computed in the usual way for each feature. It devolves from the absolute value of the real t-statistic among permuted values together with the ordinary absolute value computed when data were not permuted.

For Mann-Whitney, “significance” was assessed analogously; but bootstrapping of families was done 1,000 times.

For each of the two methods of assessing significance, we rejected the null hypothesis if the permutation p-value was less than .05. As we bootstrapped, we counted the number of bootstrap samples that resulted in rejecting the null hypothesis. Figures 5 and 6 give the distributions of p-values for the feature total cholesterol, separately for the two tests. Table 6 is a summary of
results. We conclude that individuals from the two clusters have significantly different distributions of BMI, triglycerides and HDL, and all six measures derived from the OGTT.

4.2. Familiality of the two clusters. In ordinary population-based studies, a phenotype being familial is a necessary but not sufficient condition for it being genetic, at least in part. However, our SAPPHIRe data were collected in such a peculiar fashion that hypertension, therefore insulin resistance, being familial on the one hand, and genetic on the others, are distinct issues.

SAPPHIRe was initially a study of hypertension by linkage, with insulin resistance, crudely defined, as the intermediate phenotype. It is patently a study by sibships, even though individuals were the sampling units in forming clusters as are summarized in Table 5. However, recruitment was not done in any sense that could be described as “random”. The “proband,” the initial person of a sibship enrolled, was necessarily hypertensive by criteria that have been described. However, over more than a decade SAPPHIRe policies for recruitment changed. Policies were influenced greatly by Risch and Zhang [42], although their paper’s subject was mapping quantitative trait loci (QTLs), not association. They recommended extremely discordant
sib pairs in order to increase the power of studies, when the probability of type 1 error is fixed. Thus, after 1995 purposes of recruiting sibships, SAPPHIRe policy became an attempt to recruit hypotensive sibs of a hypertensive proband. Obviously, this reduced familiality of hypertension. Since hypertension and insulin resistance, almost however defined, are related, adherence to this policy would also reduce familiality of insulin resistance. Adherence was particularly strong in Taiwan, from where a majority of SAPPHIRe Chinese female subjects were recruited. Nonetheless, it behooved us to investigate familiality of our two found clusters; that is the topic of this section.

Denote the $j^{th}$ family by $f_j$ and the cardinality of its sibship by $s_j$. Apparently $s_j$ is an integer between 1 and 6, and $1 \leq j \leq 287$. We denote people who are assigned to cluster 1 (our insulin resistant cluster) by $C_1$ and those assigned to cluster 2 by $C_2$. Familiality or lack of it is inferred by a conditional permutation test. For each cluster we compute the expected number of families that would appear in the cluster under a null hypothesis in which people are assigned to clusters at random without regard to sibship, but three observed outcomes are fixed. When the number of families is denoted by $F$, and the total number of individuals is denoted by $n$, then
Table 6
Significance tests of difference between clusters in Table 5: ratio of rejecting $H_0$.

<table>
<thead>
<tr>
<th>Ratio of rejecting $H_0$ among bootstrap samples</th>
<th>Permutation test (100 bootstraps)</th>
<th>Mann-Whitney-Wilcoxon test (1000 bootstraps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.58</td>
<td>0.61</td>
</tr>
<tr>
<td>BMI</td>
<td>0.99</td>
<td>1.0</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Total cholesterol</td>
<td>0.90</td>
<td>0.86</td>
</tr>
<tr>
<td>HDL</td>
<td>1.0</td>
<td>0.99</td>
</tr>
<tr>
<td>OGGT glucose t=0</td>
<td>0.99</td>
<td>1.0</td>
</tr>
<tr>
<td>OGGT glucose t=60</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>OGGT glucose t=120</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>OGGT insulin t=0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>OGGT insulin t=60</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>OGGT insulin t=120</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

$n = \sum_{j=1}^{F} s_j = 557$. Count the number of families that appear in $C_1$:

$$NF(C_1, k) = | \{ f_j : s_j = k \text{ and any sib of } f_j \text{ belongs to } C_1 \}|$$

(12)

$$= \sum_{j=1, s_j = k}^{F} I(f_j \cap C_1 \neq 0),$$

where $1 \leq k \leq 6$.

Our test is conditional upon $|C_1| = N_1$ (say); $|C_2| = N_2$; and $\{s_j\}$. The null hypothesis, $H_0$ is that given the three cited conditions, people are assigned to clusters at random; and the (conditional) expected numbers of families of respective sizes $j$ (2 through 6) are as if so assigned. One computes the conditional expectation:

$$E_{H_0}(|\{ f_j : s_j \geq 2 \text{ and any sib of } f_j \text{ belongs to } C_1 \}| || N_1, N_2, \{s_j\})$$

(13)

$$= \sum_{j=1, s_j \geq 2}^{F} s_j \left( 1 - \left( \frac{n - s_j}{N_1} \right) \left( \frac{n}{N_1} \right) \right)$$

Table 7 gives sibship sizes (when it is at least two), total numbers of families, observed numbers of families of respective sizes, and numbers expected under $H_0$. One needs no statistician to infer that by any reasonable test statistic, our null hypothesis is “accepted.” The third and fourth columns of Table 7 pass what the late Leonard J. Savage called “the traumatic intraocular test.”
Table 7
Sibship size and expected number of families in insulin resistant cluster under $H_0$

<table>
<thead>
<tr>
<th>Sibship size</th>
<th>num of families</th>
<th>$NF(C_1,k)$</th>
<th>expected num of families in $C_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>120</td>
<td>61</td>
<td>66.38</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>30</td>
<td>28.07</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>14</td>
<td>12.02</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4.34</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0.91</td>
</tr>
<tr>
<td>Total</td>
<td>181</td>
<td>110</td>
<td>111.72</td>
</tr>
</tbody>
</table>

5. Inference from SNPs. SAPPHIRE includes information from genotyping 293 SNPs in (candidate) genes that were pre-selected because abnormalities in them are thought to predispose to insulin resistance. These data are “unphased” as to chromosome, meaning that each SNP takes three values, two homozygotic and one heterozygotic. This section is about how we quantify the relationship of these SNP values and cluster membership, the goal being to find predictive SNPs. While the specific values of SNPs might include further information, here we associate only SNPs themselves with cluster membership, and that by mutual information [11].

The mutual information between two (discrete) random vectors $X$ and $Y$ is defined as

$$I(X : Y) = \sum_x \sum_y p(x,y)\ln \frac{p(x,y)}{p(x)p(y)},$$

(14)

where $p(x,y)$ is a joint probability mass function (pmf) of $X$ and $Y$, and $p(x)$ and $p(y)$ are marginal pmfs of $X$ and $Y$, respectively. Obviously if $X$ and $Y$ are independent, $I(X : Y) = 0$. Mutual information has been used often in statistical learning [25]. Mutual information is always non-negative. The more they are related, the higher $I(X,Y)$ is. Thus, we postulate that SNPs relevant to insulin resistance have higher mutual information with insulin resistance than other SNPs irrelevant to insulin resistance. Once relevant SNPs are identified, we also try to find their interaction with environmental variables (age and BMI). Since insulin resistance can be triggered by environmental factors, SNPs alone do not cause all insulin resistance; and both (relevant) SNPs and their interaction with age and BMI are used to classify people based on our definition of insulin resistance. Five state-of-the-art classification algorithms were considered to evaluate the procedure for finding relevant SNPs and their interaction with environmental variables. In using relevant SNPs to train classifiers and for obvious reasons, we tried to remove redundant SNPs. As an extreme example, consider a case where we have two
relevant SNPs that determine each other perfectly. We certainly want to use only one of them in building a classifier. In addition, having more features is not always good because classifiers may overfit data. Thus, we tried to remove redundant SNPs that can be predicted well by other representative SNPs. To this end, SNPs were clustered to see if there are groups of SNPs within which SNPs have high similarity (i.e., high mutual information). In clustering SNPs, similarity between SNPs was also measured by mutual information. Then each cluster was represented by a SNP that had the highest mutual information with insulin resistance within the same cluster.

Note that among 297 SNPs, 36 SNPs had a constant value, and 12 SNPs had too few people genotyped (fewer than 400 people were genotyped for those 12 SNPs; see Figure 7 for number of missing values per SNP). Since those 48 SNPs could not help us predict insulin resistance with current available information, we discarded them from further consideration; only the remaining 249 SNPs were used to find relevant SNPs and to remove redundant SNPs by clustering them.

For the remaining 249 SNPs, missing values were imputed by RPART [46]. Note also that individuals with more than 110 missing values of SNPs (corresponding to 38% of total SNPs) were discarded prior to imputation. They have significantly many missing values, and imputation may not be
reliable. This reduces the number of people from 557 to 485. See Figure 8 for number of missing SNPs per person.

5.1. Clustering SNPs. We removed redundant SNPs by clustering. Agglomerative clustering with average linkage [31][25] was used to cluster SNPs with the similarity quantified by mutual information. Since each SNP can take three possible values (major allele homozygote, heterozygote, and minor allele homozygote), a $3 \times 3$ table was formed for each pair of SNPs, of which examples are given in Tables 8(a) and 8(b). For example, in SAPPHIRe terminology (see Table 14 for its translation to more customary dbSNP accession for some SNPs) APOAV.S.1 and APOAV.S.4 were almost perfectly predictable from each other as is clear from Table 8(a). Table 8(a) and Table 8(b) give examples in SAPPHIRe terminology for which the former exhibiting high mutual information and the latter low mutual information.

To perform agglomerative clustering, we grew of a bottom-up-tree using (14) as a measure of similarity and continued to merge clusters until we were left with one cluster. Since SNPs belonging to the same node are similar, we represented each node of the tree (including terminal/leaf nodes) by a SNP with maximum mutual information with cluster membership among all SNPs belonging to the same node. We cut the tree; the SNPs representing
Table 8
3 × 3 tables: (a) mutual information: 1.2675, (b) mutual information: 0.0018

(a) APOAV.S.1 vs APOAV.S.4

<table>
<thead>
<tr>
<th></th>
<th>BB</th>
<th>Bb</th>
<th>bb</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>266</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Aa</td>
<td>1</td>
<td>175</td>
<td>0</td>
</tr>
<tr>
<td>aa</td>
<td>0</td>
<td>0</td>
<td>37</td>
</tr>
</tbody>
</table>

(b) LEPR.12 vs. PRKCZ.14

<table>
<thead>
<tr>
<th></th>
<th>BB</th>
<th>Bb</th>
<th>bb</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>94</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>Aa</td>
<td>59</td>
<td>32</td>
<td>0</td>
</tr>
<tr>
<td>aa</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

terminal nodes of the resulting sub-tree become of final selection of relevant SNPs. Cross-validation was employed to decide where to cut the tree, accuracy of classification being the criterion.

5.2. Interaction among predictors. We have emphasized that IR can be triggered by genes, environmental features, and their interactions. We discovered interactions by an approach that has proven useful in other contexts but that seems reasonably new to genetics [35][3]. First, a binary decision tree was grown using RPART [46], an R freeware version of CART [7]. We did not condition splits on any linear combinations of features. Features were SNPs selected as has been described and the three scaled functions of age and BMI (AGE², BMI², and AGE × BMI) that contain the great preponderance of information in the first two principal components of the 17 cited predictors. With CART’s equivariance to monotone transformations of coordinate axes, the squares could have been deleted, and are in results we report. The “outcome” for our two-class problem was cluster membership, with the product of empirical prior times misclassification cost taken to be equal to 1 for each class. This choice is entirely in keeping with the findings in [4] and with findings regarding support vector machines that we report here. The “no data Bayes risk” [7] is thus 1. Precise methodology for variable selection depends on each path from root node to terminal node of the tree. If, for example, a path is, “split on A, split on B, split on A,” then there are three “words” suggested: A, B, A × B if adjacent nodes are taken as suggestive of two-factor interactions, and three-factor interactions are ignored. This, then, is the basis for the three main effects and interaction chosen for inclusion in subsequent classifiers, which promise to be more accurate than...
naive CART itself. This harks back to the origins of binary tree-structured decision trees as, “automatic interaction detectors,” [45], and is particularly important in the genetic context where there are so many candidates for features available, especially interactions. Note that for any word selected for inclusion in subsequent analyses, necessarily all “sub-words” are included.

Figure 9 gives the RPART tree grown on all the data, with customary 10-fold cross-validatory choice of pruning parameter. Table 15 gives three interactions suggested by the tree.

![CART and interaction terms: BMI² and AGE² are scaled and samples satisfying split condition at each node go to the left branch.]

5.3. Prediction of Insulin Resistance based on SNPs and other features. For present purposes, the definition of IR for each person was taken to
be cluster membership, Cluster 1 being the IR group and Cluster 2 the “normal” group. We clustered SNPs to find those relevant to prediction, while removing from our list redundant SNPs, always on the basis of mutual information as described.

We tried to find interactions among SNPs and the three cited functions of age and BMI. In this section we summarize briefly the results of 10-fold cross-validation for our two-class problem. The validation included the entire process of finding relevant SNPs and interaction terms; we applied five state-of-the-art algorithms. They were support vector machines (SVM) [9][48], L1 (Lasso) regularization path algorithm for generalized linear models [38], and logistic regression with L2 penalty [39], FlexTree [29], and random forests [6]. Tables 10 - 12 summarize the accuracy of performance. For each training set in cross-validation, we selected SNPs and interaction terms, applied trained models to the cross-validatory test set to evaluate performance. We considered various losses to weight misclassification costs. In the end, we preferred the choice for which the product of empirical priors times misclassification cost were roughly equal by class. Three increasing (by inclusion) sets of features were used for each algorithm. The first consisted only of the three cited functions of age and BMI; the second included these three features plus the main effects of chosen “relevant” SNPs. The third consisted of this second set supplemented by interactions as gleaned from the tree displayed in Figure 9. Adding interactions improved performance further. Paired t-tests done for each cross-validatory test set yielded t-statistics that summarize overall accuracy when the three sets of features were compared. For the first versus the second, $t = 0.939$; for the second and the third, $t = 1.22$; while for the first versus the third, $t = 2.368$, suggesting improvement that we explore further in what follows. Table 9 summarizes 10 fold cross-validation classification for SVM with loss function 2.2:1, for which three feature sets have good overall performance without much skew in sensitivity or specificity. Since we tried to remove the effects of age and BMI in defining IR, we expected poor classification using only the three functions of age and BMI.

Note from Figure 10, where cluster membership has been projected on the (age, BMI) plane, the obvious difficulty separating our two defined classes on the basis of age and BMI alone.

Random forests and FlexTree do not perform comparably to the aforementioned three algorithms, and their performances are not reported here.
Overall classification for three feature sets: SVM with loss 2.2:1. With this loss, misclassification cost of insulin resistant class is 2.2 times higher than that of non-insulin resistant class. Products of empirical priors times misclassification are nearly equal here.

<table>
<thead>
<tr>
<th>Cross-validation Set</th>
<th>FeatureSet1</th>
<th>FeatureSet2</th>
<th>FeatureSet3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.625</td>
<td>0.6875</td>
<td>0.6667</td>
</tr>
<tr>
<td>2</td>
<td>0.4792</td>
<td>0.5208</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>0.5625</td>
<td>0.6042</td>
<td>0.6458</td>
</tr>
<tr>
<td>4</td>
<td>0.7292</td>
<td>0.5883</td>
<td>0.7083</td>
</tr>
<tr>
<td>5</td>
<td>0.6667</td>
<td>0.7292</td>
<td>0.6875</td>
</tr>
<tr>
<td>6</td>
<td>0.6667</td>
<td>0.7292</td>
<td>0.75</td>
</tr>
<tr>
<td>7</td>
<td>0.5625</td>
<td>0.7083</td>
<td>0.7083</td>
</tr>
<tr>
<td>8</td>
<td>0.5833</td>
<td>0.6667</td>
<td>0.6875</td>
</tr>
<tr>
<td>9</td>
<td>0.5417</td>
<td>0.5833</td>
<td>0.5833</td>
</tr>
<tr>
<td>10</td>
<td>0.5833</td>
<td>0.4583</td>
<td>0.5208</td>
</tr>
</tbody>
</table>

AVG 0.6 0.627 0.646

Fig 10. Clear overlapping of clusters in Age and BMI: (insulin resistant people (o)) and non-insulin resistant people (x)

5.4. Significance of SNPs and interactions. We computed a chi-square statistic to help decide whether classification by SVM and by cluster membership are unrelated. We chose SVM classification using the third of the increasing nested sets of features that were described using a a ratio of mis-
classification costs 2.1:1 for which SVM shows its best performance in terms of overall classification and renders empirical priors multiplied by respective misclassification costs about equal. The choice of misclassification renders the respective products with empirical priors almost equal by class (that is, by cluster). Computing from data in Table 13 under a hypothesis of lack of association, chi-square is

\[
\chi^2 = \sum_{i,j} \frac{(observed_{i,j} - expected_{i,j})^2}{expected_{i,j}} = 21.2.
\]

Of course, this is an example of a “maximally selected chi-square statistics,” and our cutoff is not linear in any test of features, as would be required.
Table 11
L1 regularization path algorithm: 10 fold cross-validation

(a) Age+BMI

<table>
<thead>
<tr>
<th>Loss</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Overall</th>
<th>Miscost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.8:1</td>
<td>0.473</td>
<td>0.741</td>
<td>0.648</td>
<td>232.0</td>
</tr>
<tr>
<td>1.9:1</td>
<td>0.520</td>
<td>0.717</td>
<td>0.646</td>
<td>234.1</td>
</tr>
<tr>
<td>2.0:1</td>
<td>0.532</td>
<td>0.692</td>
<td>0.631</td>
<td>245.9</td>
</tr>
<tr>
<td>2.1:1</td>
<td>0.590</td>
<td>0.619</td>
<td>0.604</td>
<td>257.7</td>
</tr>
<tr>
<td>2.2:1</td>
<td>0.610</td>
<td>0.539</td>
<td>0.579</td>
<td>283.1</td>
</tr>
<tr>
<td>2.3:1</td>
<td>0.719</td>
<td>0.491</td>
<td>0.554</td>
<td>265.7</td>
</tr>
</tbody>
</table>

(b) Age+BMI+SNPs

<table>
<thead>
<tr>
<th>Loss</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Overall</th>
<th>Miscost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.8:1</td>
<td>0.514</td>
<td>0.714</td>
<td>0.646</td>
<td>229.3</td>
</tr>
<tr>
<td>1.9:1</td>
<td>0.543</td>
<td>0.685</td>
<td>0.638</td>
<td>237.5</td>
</tr>
<tr>
<td>2.0:1</td>
<td>0.555</td>
<td>0.667</td>
<td>0.629</td>
<td>246.7</td>
</tr>
<tr>
<td>2.1:1</td>
<td>0.554</td>
<td>0.668</td>
<td>0.629</td>
<td>254.0</td>
</tr>
<tr>
<td>2.2:1</td>
<td>0.600</td>
<td>0.623</td>
<td>0.617</td>
<td>258.2</td>
</tr>
<tr>
<td>2.3:1</td>
<td>0.613</td>
<td>0.608</td>
<td>0.608</td>
<td>265.9</td>
</tr>
</tbody>
</table>

(c) Age+BMI+SNPs+Interaction terms

<table>
<thead>
<tr>
<th>Loss</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Overall</th>
<th>Miscost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.8:1</td>
<td>0.360</td>
<td>0.851</td>
<td>0.685</td>
<td>228.0</td>
</tr>
<tr>
<td>1.9:1</td>
<td>0.514</td>
<td>0.710</td>
<td>0.644</td>
<td>238.1</td>
</tr>
<tr>
<td>2.0:1</td>
<td>0.531</td>
<td>0.680</td>
<td>0.629</td>
<td>250.1</td>
</tr>
<tr>
<td>2.1:1</td>
<td>0.571</td>
<td>0.652</td>
<td>0.623</td>
<td>253.5</td>
</tr>
<tr>
<td>2.2:1</td>
<td>0.565</td>
<td>0.649</td>
<td>0.619</td>
<td>263.2</td>
</tr>
<tr>
<td>2.3:1</td>
<td>0.610</td>
<td>0.603</td>
<td>0.602</td>
<td>268.3</td>
</tr>
</tbody>
</table>

for application of the argument of [36]. However, comparison of 21.2 with for example, Table 2 of that paper suggests that the success of our classification does not owe to chance.

5.4.1. Permutation test for SNPs and interactions. We checked whether adding SNPs and interactions to age and BMI predicts cluster membership by randomly permuting SNPs across people. We performed 1,000 permutations (including the original un-permuted data) and measured area under a conventional ROC curve (AUC) for each permuted set. Then, we computed a p-value as follows:

\[
p = \frac{\text{number of permuted sets with } \text{AUC} \geq \text{AUC(SNPs)}}{1000},
\]
Table 12
Logistic regression with L2 penalty: 10 fold cross-validation

(a) Age+BMI

<table>
<thead>
<tr>
<th>Loss</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Overall</th>
<th>Miscost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.8:1</td>
<td>0.459</td>
<td>0.744</td>
<td>0.646</td>
<td>235.0</td>
</tr>
<tr>
<td>1.9:1</td>
<td>0.514</td>
<td>0.707</td>
<td>0.638</td>
<td>238.8</td>
</tr>
<tr>
<td>2.0:1</td>
<td>0.557</td>
<td>0.658</td>
<td>0.612</td>
<td>249.2</td>
</tr>
<tr>
<td>2.1:1</td>
<td>0.590</td>
<td>0.622</td>
<td>0.606</td>
<td>256.6</td>
</tr>
<tr>
<td>2.2:1</td>
<td>0.604</td>
<td>0.576</td>
<td>0.579</td>
<td>273.3</td>
</tr>
<tr>
<td>2.3:1</td>
<td>0.653</td>
<td>0.552</td>
<td>0.577</td>
<td>269.6</td>
</tr>
</tbody>
</table>

(b) Age+BMI+SNPs

<table>
<thead>
<tr>
<th>Loss</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Overall</th>
<th>Miscost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.8:1</td>
<td>0.506</td>
<td>0.719</td>
<td>0.646</td>
<td>229.7</td>
</tr>
<tr>
<td>1.9:1</td>
<td>0.550</td>
<td>0.700</td>
<td>0.648</td>
<td>230.5</td>
</tr>
<tr>
<td>2.0:1</td>
<td>0.538</td>
<td>0.674</td>
<td>0.627</td>
<td>249.8</td>
</tr>
<tr>
<td>2.1:1</td>
<td>0.553</td>
<td>0.649</td>
<td>0.615</td>
<td>260.1</td>
</tr>
<tr>
<td>2.2:1</td>
<td>0.558</td>
<td>0.639</td>
<td>0.610</td>
<td>268.6</td>
</tr>
<tr>
<td>2.3:1</td>
<td>0.572</td>
<td>0.630</td>
<td>0.608</td>
<td>273.3</td>
</tr>
</tbody>
</table>

(c) Age+BMI+SNPs+Interaction terms

<table>
<thead>
<tr>
<th>Loss</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Overall</th>
<th>Miscost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.8:1</td>
<td>0.514</td>
<td>0.734</td>
<td>0.658</td>
<td>222.8</td>
</tr>
<tr>
<td>1.9:1</td>
<td>0.521</td>
<td>0.704</td>
<td>0.642</td>
<td>239.9</td>
</tr>
<tr>
<td>2.0:1</td>
<td>0.523</td>
<td>0.683</td>
<td>0.640</td>
<td>251.6</td>
</tr>
<tr>
<td>2.1:1</td>
<td>0.545</td>
<td>0.671</td>
<td>0.627</td>
<td>255.7</td>
</tr>
<tr>
<td>2.2:1</td>
<td>0.585</td>
<td>0.671</td>
<td>0.640</td>
<td>249.0</td>
</tr>
<tr>
<td>2.3:1</td>
<td>0.600</td>
<td>0.639</td>
<td>0.623</td>
<td>260.3</td>
</tr>
</tbody>
</table>

where AUC(SNPs) is the AUC for un-permuted data.

We performed the permutation test for SNPs and interactions similarly. Since SNPs were permuted, interaction terms were identified based on permuted SNPs to measure their impact of interactions. Thus, we discovered interaction terms in each permuted set separately. The p-value was computed as in (15).

Figures 11 (a) and (b) show sensitivity and specificity of permutation tests for SNPs and SNPs and interactions, respectively; p-values for SNPs and SNPs and interactions are 0.223 and 0.028, respectively. We conclude that SNPs alone may not significantly improve classification, but SNPs along with interactions may. Further, we do fir better when specificity is high and sensitivity low than otherwise.
Table 13
Contingency table: classification vs. cluster membership. Numbers in parenthesis are expected values under the independence hypothesis.

<table>
<thead>
<tr>
<th></th>
<th>Insulin resistant</th>
<th>Insulin sensitive</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classified insulin resistant</td>
<td>214 (190.6)</td>
<td>67 (90.4)</td>
<td>284</td>
</tr>
<tr>
<td>Classified insulin sensitive</td>
<td>115 (138.4)</td>
<td>89 (65.6)</td>
<td>196</td>
</tr>
<tr>
<td>sum</td>
<td>329</td>
<td>156</td>
<td>485</td>
</tr>
</tbody>
</table>

5.5. Selected relevant SNPs and interaction terms. In the previous sections, cross-validation was performed to validate the entire procedure of defining insulin resistance with clustering and predicting insulin resistance based on age, BMI, SNPs and their interactions. In this section, the entire procedure is applied to the whole data set, and selected relevant SNPs and interaction terms are reported in Table 14 and Table 15, respectively. The tree is displayed in Figure 9 with BMI, not the original BMI$^2$ with which it was grown.

Table 14
Details regarding SNPs found to be predictive of insulin resistance

<table>
<thead>
<tr>
<th>SAPPHIRE terminology for predictive SNPs</th>
<th>dbSNP Accession</th>
<th>Human gene</th>
<th>Location</th>
<th>Frequency of major allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAMA4.S.2</td>
<td>rs1050348</td>
<td>LAMA4</td>
<td>6q21</td>
<td>0.82</td>
</tr>
<tr>
<td>CYP1B15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAMA4.S.17</td>
<td>rs1050353</td>
<td>LAMA4</td>
<td>6q21</td>
<td>0.66</td>
</tr>
<tr>
<td>LAMA4.S.22</td>
<td>rs12208872</td>
<td>LAMA4</td>
<td>6q21</td>
<td>0.66</td>
</tr>
<tr>
<td>LAMA4.S.18</td>
<td>rs3734289</td>
<td>LAMA4</td>
<td>6q21</td>
<td>0.66</td>
</tr>
<tr>
<td>FOX01A.S.4</td>
<td>rs3751437</td>
<td>FOX01</td>
<td>13q14.q</td>
<td>0.91</td>
</tr>
<tr>
<td>APOAV.S.6</td>
<td>rs662799</td>
<td>APOAV</td>
<td>11q23</td>
<td>0.74</td>
</tr>
<tr>
<td>APOAV.S.1</td>
<td>rs2072560</td>
<td>APOAV</td>
<td>11q23</td>
<td>0.74</td>
</tr>
<tr>
<td>SLC2A4.S.1</td>
<td>rs5435</td>
<td>SLC2A4</td>
<td>17q13</td>
<td>0.7</td>
</tr>
<tr>
<td>HUT2SNP5</td>
<td>rs1123617</td>
<td>HUT2</td>
<td>16q21</td>
<td>0.68</td>
</tr>
<tr>
<td>PRKCI.2</td>
<td>rs55683301</td>
<td>PRKCI</td>
<td>3q26.3</td>
<td>0.93</td>
</tr>
<tr>
<td>CD36.1</td>
<td>rs1405747</td>
<td>CD36</td>
<td>7q11.2</td>
<td>0.5</td>
</tr>
<tr>
<td>CD36.3</td>
<td>rs3211956</td>
<td>CD36</td>
<td>7q1.2</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 15
Interaction terms suggested by the Tree in Figure 9

1. BMI \times LAMA4.S.2
2. BMI \times SLC2A4.S.1
6. Summary: Insulin resistance and clustering. A novel approach to identify individuals with insulin resistance was presented in the hope that this approach will facilitate the discovery of novel genetic determinants of insulin resistance in the form of SNPs. We chose first to regress out the effects
of established determinants of insulin resistance from nine serum measures that are used commonly to quantify insulin sensitivity. Then a cluster analysis was performed on the residuals of these linearly fitted models. It produced two distinct clusters based on the means of these nine measures, an "insulin resistant" and an "insulin sensitive" cluster. To respect family structures in our data, a bootstrap method was used to sample data. Various model selection algorithms were used to estimate the number of clusters. Once insulin resistance was defined, five classification algorithms were applied to the data. We were able to improve the identification of insulin resistance individuals beyond what was possible with established determinants alone by adding SNPs and interactions.

One notes from Figure 9 and Table 15 that two genes that proved particularly important to quantifying insulin resistance were identified: LAMA4 and SLC2A4. The laminins are a family of extracellular matrix glycoproteins. They are the major noncollagenous constituent of basement membranes and have been implicated in many biological processes. Each laminin chain is a multidomain protein encoded by a distinct gene. LAMA4 is one of them. SLC2A4 encodes a protein that functions as an insulin-regulated facilitative glucose transporter. Mutations of this gene have been associated with type 2 diabetes, therefore with insulin resistance.

Acknowledgements. Authors acknowledge helpful advice regarding Gauss mixture vector quantization from Robert M. Gray, and also advice regarding our difficult job of validation from Robert Tibshirani. We acknowledge, too, conversations with SAPPHIRe colleagues over many years.
REFERENCES


imsart-aoas ver. 2009/05/21 file: InsulinResistance.tex date: August 3, 2009


INSULIN RESISTANCE: REGRESSION AND CLUSTERING


Sangho Yoon
Google
1600 Amphitheatre Parkway,
Mountain View, CA 94043, USA
E-mail: shyoon@google.com

Thomas Quertermous
Division of Cardiovascular Medicine
Department of Medicine
Falk Cardiovascular Research Center
Stanford, CA 94305, USA
E-mail: tomq1@stanford.edu

Lee-Ming Chuang
Graduate Institute of Clinical Medicine
National Taiwan University
No. 1, Jen-Ai Road, Section 1
Taipei, 100 Taiwan
E-mail: leeming@ntu.edu.tw

Richard A. Olshen
Departments of Health Research and Policy,
Electrical Engineering and Statistics
Sequoia Hall
Stanford, CA 94305, USA
E-mail: olshen@stanford.edu

Chin-Fu Hsiao
Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences
National Health Research Institutes
No. 35, Keyan Road, Zhinan Town,
Miaoli County, 350 Taiwan
E-mail: chinfu@nhri.org.tw

Chi-Min Hwu
Faculty of Medicine
National Yang-Ming University
School of Medicine
No.155, Sec.2, Linong Street
Taipei, 112 Taiwan
E-mail: chhwu@vghtpe.gov.tw

Themistocles L. Assimes
Division of Cardiovascular Medicine
Department of Medicine
Falk Cardiovascular Research Center
Stanford, CA 94305, USA
E-mail: tassimes@stanford.edu

Sangho Yoon
Google
1600 Amphitheatre Parkway,
Mountain View, CA 94043, USA
E-mail: shyoon@google.com

Thomas Quertermous
Division of Cardiovascular Medicine
Department of Medicine
Falk Cardiovascular Research Center
Stanford, CA 94305, USA
E-mail: tomq1@stanford.edu

Lee-Ming Chuang
Graduate Institute of Clinical Medicine
National Taiwan University
No. 1, Jen-Ai Road, Section 1
Taipei, 100 Taiwan
E-mail: leeming@ntu.edu.tw

Richard A. Olshen
Departments of Health Research and Policy,
Electrical Engineering and Statistics
Sequoia Hall
Stanford, CA 94305, USA
E-mail: olshen@stanford.edu