Published in final edited form as:

Genet Epidemiol. 2013 November ; 37(7): 715–725. doi:10.1002/gepi.21751.

A Multiclass Likelihood Ratio Approach for Genetic Risk Prediction Allowing for Phenotypic Heterogeneity

Yalu Wen and Qing Lu*

Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, Michigan

Abstract

The translation of human genome discoveries into health practice is one of the major challenges in the coming decades. The use of emerging genetic knowledge for early disease prediction, prevention, and pharmacogenetics will advance genome medicine and lead to more effective prevention/treatment strategies. For this reason, studies to assess the combined role of genetic and environmental discoveries in early disease prediction represent high priority research projects, as manifested in the multiple risk prediction studies now underway. However, the risk prediction models formed to date lack sufficient accuracy for clinical use. Converging evidence suggests that diseases with the same or similar clinical manifestations could have different pathophysiological and etiological processes. When heterogeneous subphenotypes are treated as a single entity, the effect size of predictors can be reduced substantially, leading to a low-accuracy risk prediction model. The use of more refined subphenotypes facilitates the identification of new predictors and leads to improved risk prediction models. To account for the phenotypic heterogeneity, we have developed a multiclass likelihood-ratio approach, which simultaneously determines the optimum number of subphenotype groups and builds a risk prediction model for each group. Simulation results demonstrated that the new approach had more accurate and robust performance than existing approaches under various underlying disease models. The empirical study of type II diabetes (T2D) by using data from the *Genes and Environment Initiatives* suggested heterogeneous etiology underlying obese and nonobese T2D patients. Considering phenotypic heterogeneity in the analysis leads to improved risk prediction models for both obese and nonobese T2D subjects.

Keywords

subphenotype; type II diabetes; area under the ROC curve

Introduction

The revolution in biotechnologies has made it possible to rapidly genotype millions of single nucleotide polymorphisms (SNPs) at an affordable cost. Benefitting from the new technologies, genome-wide association studies (GWASs) have become commonplace in biology. GWAS have shown great promise in identifying novel genetic variants that contribute to complex human diseases, such as type II diabetes (T2D) [Kilpelainen et al., 2008; Rung et al., 2009; Saxena et al., 2007; Voight et al., 2010; Wellcome Trust Case Control Consortium, 2007; Zeggini et al., 2007, 2008], cardiovascular diseases [Cluett et al.,

2009; Ellis et al., 2010; Lanktree et al., 2008; Preuss et al., 2010; Qi et al., 2011; Saleheen et al., 2010; Schaefer et al., 2009; Silander et al., 2009; Yamada et al., 2008], and bipolar disorders [Baum et al., 2008; Ferreira et al., 2008; Moskvina et al., 2009; Sklar et al., 2008]. With the increased number of findings resulted from the ongoing genetic studies, the translation of these discoveries into health practice has emerged as one of the major challenges of the coming decades. It is anticipated that the use of converging genetic knowledge for early disease prediction, prevention, and pharmacogenetics will advance genome medicine and lead to more effective personalized prevention/treatment strategies. Despite such promise, the genetic risk prediction models formed to date have lacked sufficient accuracy for clinical use. Although future research may identify additional genetic variants, such as gene-gene interactions, rare variants, and copy number variations, to further improve risk prediction of complex diseases, the effect of phenotypic heterogeneity on genetic risk prediction has been greatly overlooked [Morris et al., 2010; Warde-Farley et al., 2012].

Converging evidence suggests that subphenotypes of diseases (e.g., bladder cancer) with the same or similar clinical manifestations could have heterogeneous pathophysiological and etiological mechanisms [Blennow et al., 2006; Dror and Avraham 2009; Guey et al., 2010; McClellan and King 2010; Waring and Rosenberg 2008]. Certain disease-susceptibility variants have only been found to be associated with a specific subphenotype [Guey et al., 2010]. For example, a recent GWAS of T2D categorized cases by body mass index (BMI), and found heterogeneous genetic effects of *TCF7L2* and *FTO* between obese and nonobese forms of T2D, as well as revealed novel genes (e.g., *CAND1*) predisposing to obese T2D [Morris et al., 2010]. Indeed, the inherited predispositions to most of cancers and spectrum disorders are heterogeneous in nature, as the biological pathways can be modulated by various genetic factors, leading to the development of diseases [McClellan and King 2010; Warde-Farley et al., 2012]. Nevertheless, most existing risk prediction approaches assume the disease under investigation as one unified phenotype with homogeneous genetic/environmental causes. In the presence of phenotypic heterogeneity, they will likely attenuate the predictive power of risk predictors associated with a specific subphenotype, leading to a risk prediction model with low accuracy. Subphenotypes of a disease, defined based upon clinical symptoms, severity of illness, age at onset, and progression profile, may help form a more homogenous subphenotype group. It is worth noting that the definition of subphenotype in this paper may be broader than the traditional definition of subphenotype. We considered each subphenotype group as a subset of subjects clustered according to certain information, which includes but not limited to clinical symptoms. The use of the homogenous subphenotype group facilitates the identification of genetic variants [Lee et al., 2011; Zaitlen et al., 2012], and further improves the accuracy of risk prediction models [Morris et al., 2010]. One approach to account for phenotypic heterogeneity is through a subset analysis (i.e., separately analyzing each subphenotype). However, the disadvantage of subset analysis is that, in many cases, dividing disease samples into subphenotypes results in a much smaller sample size, thus reducing the accuracy and precision of risk prediction models. Alternatively, we could use a predefined cutoff to determine the subphenotype groups (e.g., T2D onset at less than 50 years of age). Nevertheless, such an analytical strategy does not allow us to further refine subphenotype groups to reflect the underlying homogeneous subphenotype groups. Because the subphenotype group definition relies on prior knowledge and does not take into account evidence from the data, it is possible that the defined subphenotype groups are quite different from the underlying subphenotype groups in the study sample, which could lead to low-accuracy risk prediction models.

To address this issue, we have developed a multiclass likelihood ratio (MLR) approach to iteratively determine homogenous subphenotype groups, and build a parsimonious risk prediction model for each subphenotype group. The approach is based on a forward

selection algorithm that is capable of capturing high order gene-gene/gene-environment interactions, and is applicable to high-dimensional data [Lu et al., 2012]. The proposed method allows us to start with more refined subphenotypes (e.g., T2D onset from 46 to 50 years of age), and then progressively search for homogenous subphenotype groups. Such strategy is data adaptive in the sense that it combines both prior knowledge and information from the data to determine the underlying homogenous subphenotype groups. Therefore, it provides a flexible framework to model the complex relationship between disease outcomes and predictors, allowing for heterogeneous genetic/environmental causes of a disease. In the following sections, we first lay out the details of the MLR approach, and then evaluate the performance of the proposed method through simulation studies. Finally, we conduct an empirical study by applying the new approach to the T2D dataset from the *Genes and Environment Initiatives in T2D (GEI-T2D)*.

Methods

Consider a case-control study of N individuals and L genetic and environmental predictors, where cases can be further categorized into K subphenotypes. Let $Z = (Z_1, Z_2, \dots, Z_L)$ be the measurement of L predictors and Y be the response measurement, where $Y = 0$ for controls and $Y = k, k = 1, 2, \dots, K$ for cases with k th subphenotype. In the MLR approach, we start with all K subphenotypes, treating each subphenotype as a distinct group, and then gradually cluster the subphenotypes into homogenous groups for most parsimonious risk prediction models. The general procedure of the method is demonstrated in Figure 1, and an illustrating example is provided in Appendix A.

In step 1 of MLR, we treat each subphenotype as a disease group, and form K risk prediction models using a forward selection algorithm [Lu et al., 2012; Ye et al., 2011]. Suppose for the k th subphenotype's prediction model, m_k SNPs have been selected, which form C_k multilocus genotype groups, $G_{1,k}, G_{2,k}, \dots, G_{C_k,k}$. The accuracy of the risk prediction model for each subphenotype can then be estimated using the area under the receiver operating characteristic curve (AUC),

$$AUC_{kk} = \sum_{i=1}^{R_k} \sum_{j=1}^{R_D} N_{G_{i,k}} N_{G_{j,k}} \psi[LR(G_{i,k}), LR(G_{j,k})] / \left(\sum_{i=1}^{R_k} N_{G_{i,k}} \sum_{j=1}^{R_D} N_{G_{j,k}} \right), \quad (1)$$

where $N_{G_{i,k}} (N_{G_{j,k}})$ and $R_k (R_D)$ denote the number of persons carrying risk group $G_{i,k} (G_{j,k})$ and the total number of risk groups in subphenotype k (control), respectively. The likelihood ratio (LR), which measures the risk of predictors with the subphenotype k , is defined as

$$LR(G_{i,k}) = \frac{P(G_{i,k}|Y=k)}{P(G_{i,k}|Y=0)}, \text{ and } \psi \text{ is a kernel function of the following form}$$

$$\psi[LR(G_i), LR(G_j)] = \begin{cases} 1 & \text{if } LR(G_j) < LR(G_i) \\ 0.5 & \text{if } LR(G_j) = LR(G_i) \\ 0 & \text{if } LR(G_j) > LR(G_i) \end{cases}.$$

Note that, if we treat LR as one-dimension score measuring the risk of an individual with subphenotype k , equation (1) is equivalent to the conventional nonparametric estimation of AUC [Pepe, 2003]. The AUC calculated from equation (1) can also be interpreted as the probability that a randomly chosen diseased individual has a higher LR than a randomly

chosen nondisease individual [Hanley and McNeil, 1982]. An illustrating example of *AUC* calculation is given in Appendix B.

Let $G'_{kk'}$ be the set of risk predictors for either subphenotype k or subphenotype k' (i.e., $G'_{kk'} = \{G_{i,k} \cup G_{j,k'} : i \in \{1, 2, \dots, C_k\}, j \in \{1, 2, \dots, C_{k'}\}\}$) and $LR'(G_i) = \frac{P(G_i|Y=k)}{P(G_i|Y=k')}$, where $G_i \in G'_{kk'}$. We further define the *AUC* between two subphenotypes, k and k' , as

$$AUC_{kk'} = \sum_{i=1}^{R_k} \sum_{j=1}^{R_{k'}} N_{G_i} N_{G_j} \psi[LR'(G_i), LR'(G_j)] / \left(\sum_{i=1}^{R_k} N_{G_i} \sum_{j=1}^{R_{k'}} N_{G_j} \right), \quad (2)$$

where $N_{G_i}(N_{G_j})$ and $R_k(R_{k'})$ denote the number of persons carrying risk group $G_i(G_j)$ and the total number of risk groups in subphenotype $k(k')$, respectively.

The overall performance of M models can be assessed by using a global *AUC*, defined as

$$AUC_T = \frac{\sum_{1 \leq k \leq k' \leq M} \omega_{k,k'} AUC_{kk'}}{\sum_{1 \leq k \leq k' \leq M} \omega_{k,k'}}, \quad (3)$$

where ω is the weight adjusting for sample size differences among the M disease groups and the control group. Here, we considered three commonly used weighting strategies: (1) $\omega_{k,k'} = 1$;

(2) $\omega_{k,k'} = m_k + m_{k'}$; and (3) $\omega_{k,k'} = \frac{\sqrt{m_k + m_{k'}}}{m_k \times m_{k'}}$, where m_k is the sample size of subphenotype k and $k = k'$ [Hand and Till, 2001; Wei et al., 2008]. When $k = k'$ the corresponding weighting strategies are (1) $\omega_{k,k} = 1$; (2) $\omega_{k,k} = m_k + m_c$; and (3)

$\omega_{k,k} = \frac{\sqrt{m_k + m_c}}{m_k \times m_c}$, where m_c is the number of controls. The overall accuracy, AUC_T , is determined by both AUC_{kk} (i.e., $k = k'$) and $AUC_{kk'}$. AUC_{kk} measures the accuracy of the k th risk prediction model, and can be obtained using equation (1). It decreases when heterogeneity is present within group k , which usually occurs when the group number is too small. $AUC_{kk'}$ measures the model's ability to differentiate the subphenotypes, and is calculated using equation (2). When the number of groups is too large, the chance of similarity between subphenotype groups increases, resulting in low $AUC_{kk'}$ values for those similar groups. Therefore, AUC_T measures the overall performance of M risk prediction models, and is used as the criteria to select the most parsimonious set of risk prediction models, with the optimal number of subphenotype groups.

Because each subphenotype in step 1 of MLR is considered as a separate group, it most likely overestimates the underlying homogenous subphenotype group. Low $AUC_{kk'}$ values are thus expected for those similar subphenotypes, resulting in a low AUC_T . To approach the underlying homogenous subphenotype groups, we gradually combine the subphenotypes together and search for the set of risk prediction models with the highest AUC_T . For a large number of subphenotypes, an exhaustive evaluation of all possible subphenotype combinations would be computationally expensive. Therefore, we introduce a computationally efficient algorithm to determine the optimal number of subphenotype groups. At each step of the algorithm, we evaluate the similarities among all possible pairs of groups and combine the two most similar subphenotype groups into one group. The similarity of two subphenotype groups can be assessed by $MSE_{k,k'}$,

$$MSE_{k,k'} = (\Delta \widehat{AUC}_{k,k'})^2 + Var(\Delta \widehat{AUC}_{k,k'}) \\ = (\widehat{AUC}_{k,k'} - \widehat{AUC}_{k',k})^2 + Var(\widehat{AUC}_{k,k'}) + Var(\widehat{AUC}_{k',k}) - 2Cov(\widehat{AUC}_{k,k'}, \widehat{AUC}_{k',k}),$$

where $\widehat{AUC}_{k,k'}$ and $\widehat{AUC}_{k',k}$ measure the ability of the k th model to predict the k' th subphenotype and the ability of the k' th model to predict the k th subphenotypes, respectively. $Cov(\widehat{AUC}_{k,k'}, \widehat{AUC}_{k',k})$ can be calculated as

$$Cov(\widehat{AUC}_{k,k'}, \widehat{AUC}_{k',k}) = \frac{V_D^{k,k'}}{N_D^{k,k'}} + \frac{V_{\bar{D}}}{N_{\bar{D}}},$$

where $N_D^{k,k'}$ denotes the total number of subjects with either subphenotype k or subphenotype k' ; $N_{\bar{D}}$ denotes the total number of control subjects. Let LR_i^k be the likelihood ratio identified by subphenotype k , we define

$$V_D^{k,k'} = \frac{1}{N_D^{k,k'}} \sum_{i=1}^{N_D^{k,k'}} \left\{ \left(N_{\bar{D}} \sum_{j=1}^{N_{\bar{D}}} \psi(LR_i^k, LR_j^k) - \widehat{AUC}_{k,k'} \right) \times \left(N_{\bar{D}} \sum_{j=1}^{N_{\bar{D}}} \psi(LR_i^{k'}, LR_j^{k'}) - \widehat{AUC}_{k',k} \right) \right\} \\ V_{\bar{D}} = \frac{1}{N_{\bar{D}}} \sum_{i=1}^{N_{\bar{D}}} \left\{ \left(N_D^{k,k'} \sum_{j=1}^{N_D^{k,k'}} \psi(LR_i^k, LR_j^k) - \widehat{AUC}_{k,k'} \right) \times \left(N_D^{k,k'} \sum_{j=1}^{N_D^{k,k'}} \psi(LR_i^{k'}, LR_j^{k'}) - \widehat{AUC}_{k',k} \right) \right\}.$$

If two subphenotypes belong to one underlying homogenous group, we would expect the risk prediction models formed on two subphenotypes to be similar, which results in a small $MSE_{k,k'}$. Based on the $MSE_{k,k'}$, at step M we combine the two most similar subphenotypes into one group, and reduce the number of groups to $K - M$. The forward selection algorithm [Lu et al., 2012; Ye et al., 2011] will then be used to form a risk prediction model for the newly formed subphenotype group, so that the overall performance of the $K - M$ risk prediction models can be assessed by AUC_T . We repeat the clustering process steps, and progressively combine similar subphenotypes into the same group. AUC_T reaches its highest value when the subphenotypes are clustered into an optimal number of homogeneous groups. The corresponding set of models is then chosen as the most parsimonious set of risk prediction models.

We use a K -fold cross-validation procedure to determine the optimum number of groups. In the cross-validation procedure, subjects in each subphenotype and the control groups are randomly divided into K subsets. $K - 1$ subsets are used as training set, while the remaining one is used as the testing set. MLR is applied to the training set to build risk prediction models, and AUC_T are calculated based on the testing set. The procedure is repeated K times to make sure each subset serves as the testing sample once. The optimum number of subphenotype groups is chosen as the group that maximizes the average AUC_T , and the corresponding set of risk prediction models is selected as the most parsimonious model.

Note that, although we illustrate MLR with categorical disease outcomes, the approach is flexible enough to consider prior biological knowledge and is able to handle ordinal disease

outcomes. With ordinal disease outcomes, instead of evaluating all pairwise $MSE_{k,k'}$, we only evaluate the $MSE_{k,k'}$ of adjacent subphenotype groups, and combine the adjacent subphenotype groups accordingly. For example, when the subphenotypes of T2D are defined based on age at onset (e.g., 36–40, 41–45, 46–50, 51–55 etc.), the proposed approach will only combine the adjacent subphenotypes (e.g., 36–40 and 41–45), but not the two distant subphenotypes (e.g., 36–40 and 51–55).

Results

Simulation Results

We conducted simulation studies to evaluate the performance of our proposed approach, and compared it with two approaches: (1) FROC-ALL: an approach treating all subphenotypes as a single unified phenotype; and (2) FROC-SUB: an approach performing subset analysis on each subphenotype. To be consistent with MLR, the same forward selection algorithm was used in FROC-ALL and FROC-SUB to search for most parsimonious risk prediction models. The first set of simulations was conducted to compare the performance of the three approaches with a varying number of distinct subphenotype groups under study. The second set of simulations compared the performance of the three approaches with varied disease models and effect sizes. In the third set of simulations, we evaluated their performance by considering a different degree of similarity between two subphenotypes.

Simulation I—In this simulation, we simply considered five subphenotypes with a varying number of underlying subphenotype groups under study. We started with a scenario where the five subphenotypes were mutually independent, and then considered scenarios where some of the subphenotypes shared the same genetic causes. Finally, we simulated a scenario where the underlying genetic causes of the five subphenotypes were virtually the same. For each subphenotype, we considered a simple disease model, which comprises a two-locus multiplicative interaction and an independent additive-effect locus. The effect sizes of the interaction and disease-susceptibility locus for each subphenotype were summarized in supplementary Table S1. The odds ratio of the multiplicative interaction for each subphenotype ranges from 2.0 to 2.1 and the odds ratio of the additive-effect locus for each subphenotype ranges from 2.0 to 2.2. In addition, we introduced five noise loci with minor allele frequencies ranging from 0.1 to 0.5. The genetic markers were simulated under an assumption of the Hardy-Weinberg equilibrium. For each scenario, we first simulated a general population with one million individuals. To account for the effect of sample size on the three analytical approaches, for each replicate instead of sampling the same number of individuals per subphenotype, we randomly selected 300 controls, 150 cases for each of two subphenotypes and 300 cases for each of the rest three subphenotypes. The simulation was repeated 1,000 times for each of the scenarios, and the global AUC , AUC_T , was calculated for MLR, FROC-ALL, and FROC-SUB.

The estimate of AUC_T , its bias and the mean squared error (MSE) of the three approaches were summarized in Table 1. As shown in Table 1, MLR had a better or at least comparable performance as compared to FROC-ALL and FROC-SUB. In scenarios where all of the subphenotypes had the same or a distinct disease etiology, MLR had a comparable performance with FROC-ALL or FROC-SUB. Although MLR demonstrated its robustness against a variety of underlying subphenotype combinations, the performances of FROC-SUB and FROC-ALL were highly dependent on the underlying disease mechanism. When a large proportion of subphenotypes were correlated, FROC-ALL had improved performance while FROC-SUB had poor performance. Contrarily, if the majority of the subphenotypes were unrelated, the performance of FROC-SUB improved while that of FROC-ALL decreased. The results demonstrated that properly combining subphenotypes helped to

increase sample size and predictive power, especially in a case where some of the subphenotypes were rare. We further evaluated MLR's ability of correctly combining subphenotypes. As shown in Table 1, MLR's chance of correctly combining the subphenotypes that shared the same disease etiology had an average value of 95%.

Simulation II—The complexity of the underlying disease models may also impact the performance of the new approach. In the second simulation, we simulated various disease models, and compared MLR with FROC-ALL and FROC-SUB. Similar to simulation I, we considered five subphenotypes, among which three belonged to one disease group and the remaining two belonged to the other disease group. We started with the simplest model, where each of two disease groups was associated with a unique three-locus disease model, comprised a two-locus multiplicative interaction and an independent locus with additive effect. We then increased the model's complexity by introducing two similar five-locus disease models, comprised one two-locus multiplicative interaction, one two-locus threshold interaction, and an independent locus. In the third model, we simulated two unique three-locus disease models, comprised one two-locus multiplicative interaction and an independent locus. Finally, we introduced two very different models for the two diseases groups. One disease model included a two-locus multiplicative interaction and an independent locus with additive effect, while the other comprised one two-locus multiplicative interaction, one two-locus threshold interaction, and an independent locus. The details of the simulation settings were summarized in supplementary Table S2. Similar to the first simulation, we also included five noise loci.

As shown in Table 2, the performance of all three approaches decreased with an increase in the model's complexity. This can be explained by the fact that as the complexity of the models increases, it is less likely that the forward selection algorithm can identify all of the risk loci with small to moderate effect sizes, leading to the reduced predictive power of all three approaches. Nevertheless, among all of the simulated scenarios, the MLR approach outperformed the other two approaches. Same as the first simulation, we also calculated the probability of wrongly combining subphenotype groups. As shown in Table 2, while the effect size did not have a significant impact on the performance of the approach, the complexity of the model did affect MLR's ability to correctly combine subphenotypes. In general, the less complex the model is, the more likely MLR would combine subphenotypes correctly. Nevertheless, even for the most complex simulated model, there was an average of 86% chance that the MLR approach would correctly combine the subphenotypes.

Simulation III—In this simulation, we mimicked real disease scenarios, where subphenotypes may not have the exact same disease etiologies but share a certain proportion of causal loci. For simplicity, we only considered two subphenotypes with a two-locus multiplicative interaction and an independent locus with additive effect. We started with a disease model in which the two subphenotypes were independent of each other, and then gradually increased the shared genetic components, until the two subphenotypes shared the same causal loci. For all models, five noise loci were introduced. The details of this set of simulations were summarized in supplementary Table S3.

As illustrated in Table 3, MLR had comparable or better performance as compared to the other two approaches. In an extreme case where the two subphenotypes were independent of each other, MLR treated them as distinct disease groups, and thus had a comparable performance to FROC-SUB. In another extreme case, when the two subphenotypes shared the same disease etiology, MLR combined them into one group, thus performing closely to FROC-ALL. In either case, MLR chose the best way to analyze the data without prespecifying the relationship between the subphenotypes. We considered this important, as the relationship between subphenotypes was not usually known in advance.

Real Data Application

We applied MLR to the datasets from the *GEI-T2D* study. The *GEI-T2D* study is one of the largest and most comprehensive case-control studies conducted to date, aimed at discovering genetic variants and gene-environmental interactions contributing to T2D. Study samples in *GEI-T2D* were selected from two large studies: the Nurses' Health Study and the Health Professionals Follow-up Study. The *GEI-T2D* included standardized diagnostic assessments of T2D, for which an ample number of cases ($n = 2,720$) and noncases ($n = 3,180$) exist. The *GEI-T2D*'s assessment plans for environmental conditions and personal characteristics (e.g., BMI) were also guided by standardized interview protocols and assessments, as described in prior *GEI-T2D* publications [Cornelis et al., 2009; Hu et al., 2001]. BMI is a well-established predictor for T2D, but T2D can also develop in people with normal BMI, suggesting different biology pathways among normal, overweight, and obese T2D cases. To account for potential phenotypic heterogeneity among T2D cases, we first categorized all subjects into normal weight ($\text{BMI} \leq 25$), overweight ($25 < \text{BMI} \leq 30$), and obese ($\text{BMI} > 30$) groups based on their BMI, and then applied MLR to the *GEI-T2D* dataset to build the T2D risk prediction model. Fifty-one SNPs, that have been previously reported to be associated with T2D, were considered in the analysis. The details of the selected SNPs were listed in supplementary Table S4. Among the 51 T2D-associated SNPs, 28 can be found directly from the *GEI-T2D* dataset, and the remaining SNPs were imputed using the model-based imputation software IMPUTE2, with the CEU population from the HapMap phase III and the 1000 Genomes Project as the reference panels [Howie et al., 2009; Marchini et al., 2007]. In addition to the 51 T2D-associated genetic variants, environmental factors, including smoking, alcohol consumption, physical activity, and diet, were also considered in the analysis.

Based on the *GEI-T2D* data, MLR classified normal weight and overweight T2D patients as one homogeneous sub-phenotype group, and obese T2D patients as the other subphenotype group. Four predictors, alcohol consumption, smoking, rs7901695 (*TCF7L2*), and rs6931514 (*CDKALI*), were selected into the risk prediction model for normal weight and overweight subjects, and the corresponding *AUC* reached 0.6149 (95% CI: (0.5440, 0.6758)). For obese subjects, six predictors, alcohol consumption, rs4607103 (*ADAMTS9*), rs564398 (*CDKN2A/2B*), rs5215 (*KCNJ11*), rs4506565 (*TCF7L2*), and rs8042680 (*FTO*), were selected into the final risk prediction model, and the corresponding *AUC* attained 0.6703 (95% CI: (0.5472, 0.7934)). The detailed results from MLR were summarized in Table 4. For comparison purposes, we also formed risk prediction models using FROC-SUB and FROC-ALL. The predictors selected by the FROC-SUB for normal weight and overweight subjects were similar to those identified by MLR, which indicates similar genetic mechanisms for normal weight and overweight T2D patients. The *AUC* of the predictive model for normal weight and overweight subjects from FROC-SUB were also close to that from the MLR approach, but the *AUC* from MLR had a narrower confidence interval than that from FROC-SUB due to the increased sample size. The risk prediction model using FROC-ALL only selected two predictors, alcohol consumption and smoking, and had an *AUC* value of 0.5922 (95% CI: (0.5399, 0.6444)), which was lower than those for MLR and FROC-SUB. As shown in the results, the risk predictors of normal/overweight subjects were quite different from those of obese subjects. When all of the T2D cases were lumped together to form a single disease group, the strong genetic variants predisposing to specific subphenotypes became negligible in the whole population.

Discussion

Risk prediction capitalizing on genetic and environmental information hold great promise for individualized disease prediction and prevention [Guttmacher and Collins, 2005;

Rogowski et al., 2009]. Nevertheless, for most of common complex diseases, the formed risk prediction models lack sufficient accuracy for potential clinical use [Lango et al., 2008; Meigs et al., 2008; Pharoah et al., 2008]. The low utility of predictive models may be attributed to heterogeneous etiological processes among disease subphenotypes (i.e., phenotypic heterogeneity) [Blennow et al., 2006; McClellan and King, 2010; Morris et al., 2010; Urbach et al., 2012; Waring and Rosenberg, 2008]. When disease heterogeneity is present, risk prediction studies likely yield attenuated estimates for predictors associated with special subphenotypes, leading to a predictive model with suboptimal performance [Morris et al., 2010]. To consider phenotypic heterogeneity [Blennow et al., 2006; McClellan and King, 2010; Morris et al., 2010; Urbach et al., 2012; Waring and Rosenberg, 2008] in risk prediction studies, we have developed a non-parametric approach, MLR, to simultaneously determine the cohesiveness of subphenotypes and build a risk prediction model accordingly.

An accurate subphenotype definition, which reflects the underlying clinical symptoms and molecular pathology, can help to form homogeneous case groups, leading to a more accurate risk prediction model for each case group. However, practically, there are generally no gold standards for defining homogeneous subphenotypes, especially when the clinical manifestations of diseases are indistinguishable. The MLR approach considers more refined subphenotypes, and then gradually combines the initial subphenotypes to simultaneously search for the homogenous subphenotype groups and the most parsimonious risk prediction models. Through simulations, we have demonstrated that the MLR approach could accurately detect the underlying subphenotype groups with various disease models. In the scenario that the two subphenotypes shared the same underlying disease mechanisms, MLR combined the two subphenotypes and treated them as one homogeneous phenotype group. On the other hand, when two subphenotypes were independent of each other, MLR treated them as two distinct disease subphenotypes and built predictive models for each of them. In circumstances where the clinically indistinguishable subphenotypes shared some of risk factors, the proposed approach partitioned subphenotypes depending on their degree of sharing of disease mechanisms. Through simulations, we found MLR outperformed FROC-ALL, an approach treating all subphenotypes as a single unified phenotype, in all scenarios except for the circumstance where all of the subphenotypes shared the same disease mechanism. This indicates that building a predictive model without considering phenotype heterogeneity can attenuate the accuracy of the model. Similarly, the MLR approach performed better than FROC-SUB, an approach treating each subphenotype distinct phenotype, except for the circumstance where all subphenotypes are different. This suggests that treating each subphenotype as a unique entity reduces predictive accuracy and precision when some of the subphenotypes share the same or similar disease mechanisms.

We further applied MLR to build a predictive model for T2D by using the *GEI-T2D* dataset. We hypothesized that the disease mechanisms of T2D might be different among normal weight, overweight, and obese patients, and therefore classified study subjects into three categories according to their BMI. MLR combined normal weight and overweight subjects to form one homogeneous group ($AUC = 0.6149$) and left the obese subjects to form another subphenotype group ($AUC = 0.6703$). For comparison purpose, we also applied the FROC-ALL and FROC-SUB approaches to the same data. FROC-ALL only selected smoking and alcohol consumption, and the accuracy of formed model ($AUC = 0.5922$) was considerably lower than that of MLR. This demonstrated that strong genetic variants predisposing to specific subphenotypes in a relative small homogeneous population might be negligible in the whole population. In addition, for the risk prediction model of normal weight and overweight subjects, MLR achieved a similar AUC as FROC-SUB but with a narrower confidence interval of AUC , indicating that combining sub-phenotypes sharing similar disease mechanisms may increase the precision of predictive models. Our empirical study

also confirmed previous findings of the heterogeneous effects of *FTO* and *TCF7L2* [Morris et al., 2010; Timpson et al., 2009]. The result indicated that *FTO* played an important role in obese T2D, while no evidence suggested *FTO* contributed significantly to normal weight and overweight T2D. With respect to *TCF7L2*, though it is an important predictor for both obese and not-obese subjects, the relevant importance of the gene is different for the two subphenotypes.

It is worth noting that the proposed method could be computationally intensive when a large number of genetic variants are involved (e.g., genome-wide data). This is primarily due to the exponentially increased search space with the increased number of genetic variants. Incorporating prior biology knowledge into approach (e.g., using the biology knowledge to limit the possible combinations) could further reduce the search space. Computational techniques, such as parallel computing, could be also used to boost the speed. Further development is needed to reduce the computation burden of the approach, making it applicable to genome-wide data.

Though MLR is data adaptive, it still depends on the sub-phenotypes used to define homogenous case groups. If the selected subphenotypes are not informative for phenotypic heterogeneity, it is more likely that MLR will treat all subphenotypes as one homogeneous group, resulting in no improvement of risk prediction models. Therefore, new approaches aimed at selecting appropriate subphenotypes to capture phenotype heterogeneity need to be developed. With the better defined subphenotypes and identification of new risk predictors (e.g., rare variants and copy number of variation), we expect that the accuracy of the risk predictive models could be further improved.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We appreciate critical input from two anonymous reviewers. This work was supported by the National Institute of Dental & Craniofacial Research under Award Number R03DE022379 and by the National Institute on Drug Abuse under Award Number 1K01DA033346-01A1. The datasets used for the analyses described in this manuscript were obtained from dbGaP through dbGaP accession number [phs000091].

References

- Baum AE, Hamshere M, Green E, Cichon S, Rietschel M, Nothen MM, Craddock N, McMahon FJ. Meta-analysis of two genome-wide association studies of bipolar disorder reveals important points of agreement. *Mol Psychiatry*. 2008; 13(5):466–467. [PubMed: 18421293]
- Blennow K, de Leon MJ, Zetterberg H. Alzheimer's disease. *Lancet*. 2006; 368(9533):387–403. [PubMed: 16876668]
- Cluett C, McDermott MM, Guralnik J, Ferrucci L, Bandinelli S, Miljkovic I, Zmuda JM, Li R, Tranah G, Harris T, et al. The 9p21 myocardial infarction risk allele increases risk of peripheral artery disease in older people. *Circ Cardiovasc Genet*. 2009; 2(4):347–353. [PubMed: 20031606]
- Cornelis MC, Qi L, Zhang C, Kraft P, Manson J, Cai T, Hunter DJ, Hu FB. Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. *Ann Intern Med*. 2009; 150(8):541–550. [PubMed: 19380854]
- Dror AA, Avraham KB. Hearing loss: mechanisms revealed by genetics and cell biology. *Annu Rev Genet*. 2009; 43:411–437. [PubMed: 19694516]
- Ellis KL, Pilbrow AP, Frampton CM, Doughty RN, Whalley GA, Ellis CJ, Palmer BR, Skelton L, Yandle TG, Palmer SC, et al. A common variant at chromosome 9P21.3 is associated with age of onset of coronary disease but not subsequent mortality. *Circ Cardiovasc Genet*. 2010; 3(3):286–293. [PubMed: 20400779]

- Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, Fan J, Kirov G, Perlis RH, Green EK, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet.* 2008; 40(9):1056–1058. [PubMed: 18711365]
- Guey LT, Garcia-Closas M, Murta-Nascimento C, Lloreta J, Palencia L, Kogevinas M, Rothman N, Vellalta G, Calle ML, Marenne G, et al. Genetic susceptibility to distinct bladder cancer subphenotypes. *Eur Urol.* 2010; 57(2):283–292. [PubMed: 19692168]
- Guttmacher AE, Collins FS. Realizing the promise of genomics in biomedical research. *JAMA.* 2005; 294(11):1399–1402. [PubMed: 16174701]
- Hand JD, Till JR. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach Learn.* 2001; 45(2):171–186.
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982; 143(1):29–36. [PubMed: 7063747]
- Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 2009; 5(6):e1000529. [PubMed: 19543373]
- Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG, Willett WC. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *N Engl J Med.* 2001; 345(11):790–797. [PubMed: 11556298]
- Kilpelainen TO, Lakka TA, Laaksonen DE, Lindstrom J, Eriksson JG, Valle TT, Hamalainen H, Ilanne-Parikka P, Keinänen-Kiukaanniemi S, Lindi V, et al. SNPs in PPARG associate with type 2 diabetes and interact with physical activity. *Med Sci Sports Exerc.* 2008; 40(1):25–33. [PubMed: 18091023]
- Lango H, Palmer CN, Morris AD, Zeggini E, Hattersley AT, McCarthy MI, Frayling TM, Weedon MN. Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes.* 2008; 57(11):3129–3135. [PubMed: 18591388]
- Lanktree M, Oh J, Hegele RA. Genetic testing for atherosclerosis risk: inevitability or pipe dream? *Can J Cardiol.* 2008; 24(11):851–854. [PubMed: 18987759]
- Lee PH, Bergen SE, Perlis RH, Sullivan PF, Sklar P, Smoller JW, Purcell SM. Modifiers and subtype-specific analyses in whole-genome association studies: a likelihood framework. *Hum Hered.* 2011; 72(1):10–20. [PubMed: 21849790]
- Lu Q, Wei C, Ye C, Li M, Elston RC. A likelihood ratio-based Mann-Whitney approach finds novel replicable joint gene action for type 2 diabetes. *Genet Epidemiol.* 2012; 36(6):583–593. [PubMed: 22760990]
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet.* 2007; 39(7):906–913. [PubMed: 17572673]
- McClellan J, King MC. Genetic heterogeneity in human disease. *Cell.* 2010; 141(2):210–217. [PubMed: 20403315]
- Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino RB Sr, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med.* 2008; 359(21):2208–2219. [PubMed: 19020323]
- Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, Hattersley AT, McCarthy MI. A powerful approach to sub-phenotype analysis in population-based genetic association studies. *Genet Epidemiol.* 2010; 34(4):335–343. [PubMed: 20039379]
- Moskvina V, Craddock N, Holmans P, Nikolov I, Pahwa JS, Green E, Owen MJ, O'Donovan MC. Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. *Mol Psychiatry.* 2009; 14(3):252–260. [PubMed: 19065143]
- Pepe, M. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* New York: Oxford University Press; 2003.
- Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. *N Engl J Med.* 2008; 358(26):2796–2803. [PubMed: 18579814]
- Preuss M, König IR, Thompson JR, Erdmann J, Absher D, Assimes TL, Blankenberg S, Boerwinkle E, Chen L, Cupples LA, et al. Design of the Coronary ARtery Disease Genome-Wide Replication

And Meta-Analysis (CARDIoGRAM) Study: A genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls. *Circ Cardiovasc Genet.* 2010; 3(5):475–483. [PubMed: 20923989]

- Qi L, Ma J, Qi Q, Hartiala J, Allayee H, Campos H. Genetic risk score and risk of myocardial infarction in Hispanics. *Circulation.* 2011; 123(4):374–380. [PubMed: 21242481]
- Rogowski WH, Grosse SD, Khoury MJ. Challenges of translating genetic tests into clinical and public health practice. *Nat Rev Genet.* 2009; 10(7):489–495. [PubMed: 19506575]
- Rung J, Cauchi S, Albrechtsen A, Shen L, Rocheleau G, Cavalcanti-Proenca C, Bacot F, Balkau B, Belisle A, Borch-Johnsen K, et al. Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nat Genet.* 2009; 41(10):1110–1115. [PubMed: 19734900]
- Saleheen D, Alexander M, Rasheed A, Wormser D, Soranzo N, Hammond N, Butterworth A, Zaidi M, Haycock P, Bumpstead S, et al. Association of the 9p21.3 locus with risk of first-ever myocardial infarction in Pakistanis: case-control study in South Asia and updated meta-analysis of Europeans. *Arterioscler Thromb Vasc Biol.* 2010; 30(7):1467–1473. [PubMed: 20395598]
- Saxena R, Voight BF, Lyssenko V, Burt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science.* 2007; 316(5829):1331–1336. [PubMed: 17463246]
- Schaefer AS, Richter GM, Groessner-Schreiber B, Noack B, Nothnagel M, El Mokhtari NE, Loos BG, Jepsen S, Schreiber S. Identification of a shared genetic susceptibility locus for coronary heart disease and periodontitis. *PLoS Genet.* 2009; 5(2):e1000378. [PubMed: 19214202]
- Silander K, Tang H, Myles S, Jakkula E, Timpson NJ, Cavalli-Sforza L, Peltonen L. Worldwide patterns of haplotype diversity at 9p21.3, a locus associated with type 2 diabetes and coronary heart disease. *Genome Med.* 2009; 1(5):51. [PubMed: 19463184]
- Sklar P, Smoller JW, Fan J, Ferreira MA, Perlis RH, Chambert K, Nimgaonkar VL, McQueen MB, Faraone SV, Kirby A, et al. Whole-genome association study of bipolar disorder. *Mol Psychiatry.* 2008; 13(6):558–569. [PubMed: 18317468]
- Timpson NJ, Lindgren CM, Weedon MN, Randall J, Ouwehand WH, Strachan DP, Rayner NW, Walker M, Hitman GA, Doney AS, et al. Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome-wide association data. *Diabetes.* 2009; 58(2):505–510. [PubMed: 19056611]
- Urbach D, Lupien M, Karagas MR, Moore JH. Cancer heterogeneity: origins and implications for genetic association studies. *Trends Genet.* 2012; 28(11):538–543. [PubMed: 22858414]
- Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet.* 2010; 42(7):579–589. [PubMed: 20581827]
- Warde-Farley D, Brudno M, Morris Q, Goldenberg A. Mixture model for subphenotyping in GWAS. *Pac Symp Biocomput.* 2012; 2012:363–374. [PubMed: 22174291]
- Waring SC, Rosenberg RN. Genome-wide association studies in Alzheimer disease. *Arch Neurol.* 2008; 65(3):329–334. [PubMed: 18332245]
- Wei Z, Li M, Rebbeck T, Li H. U-statistics-based tests for multiple genes in genetic association studies. *Ann Hum Genet.* 2008; 72(Pt 6):821–833. [PubMed: 18691161]
- Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007; 447(7145):661–678. [PubMed: 17554300]
- Yamada Y, Ichihara S, Nishida T. Molecular genetics of myocardial infarction. *Genomic Med.* 2008; 2(1–2):7–22. [PubMed: 18704761]
- Ye C, Cui Y, Wei C, Elston RC, Zhu J, Lu Q. A non-parametric method for building predictive genetic tests on high-dimensional data. *Hum Hered.* 2011; 71(3):161–170. [PubMed: 21778735]
- Zaitlen N, Lindstrom S, Pasaniuc B, Cornelis M, Genovese G, Pollack S, Barton A, Bickeboller H, Bowden DW, Eyre S, et al. Informed conditioning on clinical covariates increases power in case-control association studies. *PLoS Genet.* 2012; 8(11):e1003032. [PubMed: 23144628]

- Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*. 2007; 316(5829):1336–1341. [PubMed: 17463249]
- Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*. 2008; 40(5):638–645. [PubMed: 18372903]

Appendix A: An Illustrating Example of the MLR Approach

In this example, we consider three subphenotypes, Sub-1, Sub-2, and Sub-3. Among these three subphenotypes, two of them shared the same disease mechanism (i.e., Sub-1 has its own disease mechanism, while Sub-2 and Sub-3 share the same disease mechanism). We assume locus A is the risk predictor for Sub-1, and loci B and C are the risk predictors for both Sub-2 and Sub-3.

We randomly split the samples into M subsets, where $M-1$ subsets are used as training set, and the remaining one is used as the test set. A set of candidate models are built on the training set, and are then evaluated on the test set. In step 1 of MLR, we treat each of the three subphenotypes as distinct disease outcomes, and form a risk prediction model for each subphenotype using a forward selection algorithm. We assume locus A, locus B, and loci B and C are selected into the risk prediction models for Sub-1, Sub-2, and Sub-3, respectively (i.e., we assume locus C is not selected into the risk prediction model for Sub-2 because of the small number of samples with Sub-2). We then assess the overall performance of three risk prediction models in the test set using AUC_T (see equation (3)). Because each subphenotype in step 1 of MLR is considered as a separate group, it most likely overestimates the underlying homogenous subphenotype groups. In the next few steps, we gradually combine the similar subphenotypes, and search for the optimum number of subphenotype groups. In step 2, we measure the similarities between subphenotypes using $MSE_{k,k'}$, and then combine two most similar subphenotypes. Because Sub-2 and Sub-3 shared the same disease mechanisms, therefore the corresponding $MSE_{2,3}$ is smaller than $MSE_{1,2}$ and $MSE_{1,3}$ (i.e., Sub-2 and Sub-3 are more similar than the other pairs of subphenotypes). Therefore, we combine Sub-2 and Sub-3 into a subphenotype group, denoted by Sub-2-3. We apply forward selection on the newly formed subphenotype group (i.e., Sub-2-3) and form a new risk prediction model. We assume that, in step 2, both loci B and C are selected into the new risk prediction model due to the increase of sample size in the combined group. We then assess the overall performance of the two risk prediction models formed at step 2 by calculating AUC_T in the test set. In step 3, we continue combining the subphenotypes so that all three subphenotypes are combined into one group. A risk prediction model is built for the case group, and its performance is evaluated by using AUC_T in the test set.

The above process is repeated M times, with each of the M subsets used exactly once as the test set. The optimum number of subphenotype groups is chosen as the one that maximize the averaged AUC_T . In step 1, each subphenotype is considered as a distinct disease outcome, which overestimates the underlying subphenotype groups. In such case, risk predictors with moderate effect size (e.g., locus C) are likely missed due to the small sample size of subphenotype groups, which could result in a low AUC_T . In step 3, we consider three subphenotypes as one unified phenotype. Because genetic cause of Sub-1 is different from that of the other two, risk predictors associated with a specific subphenotype (e.g., locus A) may not be detected, leading to risk prediction models with low accuracy. In this particular example, combining subphenotypes into two groups gives the highest model performance, and the corresponding risk prediction models are chosen as the most parsimonious models.

Table B1

Summary statistics of the data

Risk group	Number of cases	Number of controls	Likelihood ratio (LR)
AA	40	12	$LR(AA) = \frac{P(G=AA Y=k)}{P(G=AA Y=0)} = 3.333$
Aa	20	15	$LR(Aa) = \frac{P(G=Aa Y=k)}{P(G=Aa Y=0)} = 1.333$
aa	10	23	$LR(aa) = \frac{P(G=aa Y=k)}{P(G=aa Y=0)} = 0.435$
Total	70	50	

Appendix B: A Simple Example of Equation (1)

We use a simple example to illustrate how to evaluate the accuracy of a risk prediction model using equation (1). Suppose we have a genetic risk predictor A, which has three genotypes, AA, Aa, and aa, we calculate the likelihood ratio (*LR*) of each genotype as the ratio of the genotype frequencies in cases (i.e., those with the subphenotype *k*) and in controls (Table B1). We treat *LR* as one-dimension risk score, and evaluate how well this score predict the disease outcome. For such purpose, we evaluate all possible case-control pairs (i.e., 3,500 pairs), and calculate the numbers of pairs where a case has a high *LR* than a control (i.e., 1,980 pairs) and a case has an equal *LR* with a control (i.e., 1,010 pairs). AUC from equation (1) can be obtained as the probability that a diseased individual has a higher *LR* than a control individual with the consideration of ties,

$$\begin{aligned}
 AUC_{kk} &= \frac{\sum_{i=1}^{R_k} \sum_{j=1}^{R_{\bar{D}}} N_{G_{i,k}} N_{G_{j,k}} \psi[LR(G_{i,k}), LR(G_{j,k})]}{\left(\sum_{i=1}^{R_k} N_{G_{i,k}} \sum_{j=1}^{R_{\bar{D}}} N_{G_{j,k}} \right)} \\
 &= \frac{\left\{ \begin{aligned} &40 \times [12 \times \psi(LR(AA), LR(AA)) + 15 \times \psi(LR(AA), LR(Aa)) + 23 \times \psi(LR(AA), LR(aa))] + \\ &20 \times [12 \times \psi(LR(Aa), LR(AA)) + 15 \times \psi(LR(Aa), LR(Aa)) + 23 \times \psi(LR(Aa), LR(aa))] + \\ &10 \times [12 \times \psi(LR(aa), LR(AA)) + 15 \times \psi(LR(aa), LR(Aa)) + 23 \times \psi(LR(aa), LR(aa))] \end{aligned} \right\}}{(70 \times 50)} \\
 &= \{1,980 \times 1 + 1,010 \times 0.5 + 510 \times 0\} / (70 \times 50) \\
 &= 2,485 / 3,500 = 0.71.
 \end{aligned}$$

Figure 1.
The flowchart of the MLR process.

Table 1

Mean, bias, standard deviation and mean squared error of AUC_T in Simulation I

Subphenotypes correlation ^a	Weight	AUC_T (MLR)					AUC_T (PROC-SUB)					AUC_T (PROC-ALL)				
		Mean	Bias ^b	SD ^c	MSE ^d	P ^h	Mean	Bias ^b	SD ^c	MSE ^d	P ^h	Mean	Bias ^b	SD ^c	MSE ^d	P ^h
1;2;3;4;5	Weight1 ^e	0.7066	0.0681	0.0136	0.0048	0.0025	0.7067	0.0680	0.0135	0.0048	0.0025	0.5279	0.2468	0.0192	0.0613	0.0025
	Weight2 ^f	0.7060	0.0673	0.0131	0.0047	0.0017	0.7061	0.0672	0.0130	0.0047	0.0017		0.2454		0.0606	0.0017
	Weight3 ^g	0.7074	0.0696	0.0157	0.0051	0.0100	0.7078	0.0692	0.0149	0.0050	0.0100		0.2491		0.0624	0.0100
(1,2);3;4;5	Weight1 ^e	0.6991	0.0744	0.0223	0.0060	0.0675	0.6944	0.0791	0.0138	0.0064	0.0675	0.5398	0.2337	0.0207	0.0550	0.0675
	Weight2 ^f	0.6992	0.0735	0.0208	0.0058	0.0825	0.6952	0.0776	0.0133	0.0062	0.0825		0.2329		0.0547	0.0825
	Weight3 ^g	0.6994	0.0763	0.0251	0.0065	0.0733	0.6936	0.0822	0.0153	0.0070	0.0733		0.2359		0.0561	0.0733
(1,2);3;(4,5)	Weight1 ^e	0.6869	0.0787	0.0271	0.0069	0.0833	0.6807	0.0849	0.0127	0.0074	0.0833	0.5501	0.2156	0.0207	0.0469	0.0833
	Weight2 ^f	0.6915	0.0762	0.0248	0.0064	0.0683	0.6830	0.0847	0.0122	0.0073	0.0683		0.2176		0.0478	0.0683
	Weight3 ^g	0.6820	0.0804	0.0307	0.0074	0.0916	0.6778	0.0846	0.0142	0.0074	0.0916		0.2123		0.0455	0.0916
1;2;(3,4,5)	Weight1 ^e	0.6895	0.0700	0.0228	0.0054	0.0775	0.6637	0.0957	0.0126	0.0093	0.0775	0.6003	0.1591	0.0225	0.0258	0.0775
	Weight2 ^f	0.6948	0.0688	0.0244	0.0053	0.0875	0.6629	0.1008	0.0121	0.0103	0.0875		0.1633		0.0272	0.0875
	Weight3 ^g	0.6865	0.0704	0.0222	0.0054	0.0525	0.6661	0.0908	0.0142	0.0084	0.0525		0.1566		0.0250	0.0525
(1,2);(3,4,5)	Weight1 ^e	0.6839	0.0659	0.0300	0.0052	0.0725	0.6489	0.1009	0.0124	0.0103	0.0725	0.6006	0.1492	0.0234	0.0228	0.0725
	Weight2 ^f	0.6888	0.0672	0.0298	0.0054	0.0575	0.6495	0.1065	0.0119	0.0115	0.0575		0.1554		0.0247	0.0575
	Weight3 ^g	0.6773	0.0616	0.0281	0.0046	0.0717	0.6493	0.0897	0.0138	0.0082	0.0717		0.1384		0.0197	0.0717
1;(2,3,4,5)	Weight1 ^e	0.6834	0.0664	0.0262	0.0051	0.0291	0.6209	0.1290	0.0120	0.0168	0.0291	0.6637	0.0861	0.0195	0.0078	0.0291
	Weight2 ^f	0.6956	0.0616	0.0278	0.0046	0.0400	0.6153	0.1418	0.0114	0.0202	0.0400		0.0934		0.0091	0.0400
	Weight3 ^g	0.6745	0.0668	0.0226	0.0050	0.0092	0.6315	0.1099	0.0135	0.0123	0.0092		0.0776		0.0064	0.0092
(1,2,3,4,5)	Weight1 ^e	0.6911	0.0280	0.0178	0.0011		0.5618	0.1572	0.0108	0.0248		0.6911	0.0280	0.0178	0.0011	
	Weight2 ^f	0.6911	0.0280	0.0178	0.0011		0.5670	0.1521	0.0105	0.0232			0.0280		0.0011	
	Weight3 ^g	0.6911	0.0280	0.0178	0.0011		0.5526	0.1665	0.0119	0.0278			0.0280		0.0011	

^aSubphenotypes in parentheses shared the same disease mechanisms.

^bBias = the true AUC_T – the estimated AUC_T .

^c SD = Standard deviation.

^d Mean squared error = Bias² + SD².

^e $\omega_{k,k'} = \omega_{k,k} = 1$.

^f $\omega_{k,k'} = m_k + m_{k'}$ and $\omega_{k,k} = m_k + m_c$, where m_k is the sample size of subphenotype k and m_c is the sample size of controls.

^g $\omega_{k,k'} = \frac{\sqrt{m_k + m_{k'}}}{m_k \times m_{k'}}$ and $\omega_{k,k} = \frac{\sqrt{m_k + m_c}}{m_k \times m_c}$, where m_k is the sample size of subphenotype k and m_c is the sample size of controls.

^h Probability of incorrectly combine subphenotypes.

Table 2

Mean, standard deviation, mean squared error of AUC_T in Simulation II

Models ^h	Weight	AUC_T (MLR)					AUC_T (FROC-SUB)					AUC_T (FROC-ALL)				
		Mean	Bias ^d	SD ^b	MSE ^c	Pr ^g	Mean	Bias ^d	SD ^b	MSE ^c	Mean	Bias ^d	SD ^b	MSE ^c	Mean	Bias ^d
2_2	Weight1 ^d	0.6830	0.0634	0.0183	0.0044	0.0325	0.6460	0.1362	0.0106	0.0102	0.6102	0.1003	0.0220	0.0190	0.6102	0.1003
	Weight2 ^e	0.6879	0.0647	0.0188	0.0045	0.0267	0.6460	0.1424	0.0102	0.0115		0.1066		0.0208		0.1066
	Weight3 ^f	0.6750	0.0604	0.0173	0.0039	0.0325	0.6475	0.1252	0.0118	0.0079		0.0880		0.0162		0.0880
5_5	Weight1 ^d	0.7014	0.0976	0.0398	0.0111	0.1467	0.6684	0.1839	0.0122	0.0172	0.6152	0.1306	0.0230	0.0343	0.6152	0.1306
	Weight2 ^e	0.7051	0.1006	0.0411	0.0118	0.1350	0.6692	0.1905	0.0118	0.0188		0.1365		0.0368		0.1365
	Weight3 ^f	0.6964	0.0910	0.0367	0.0096	0.1490	0.6686	0.1722	0.0136	0.0143		0.1187		0.0302		0.1187
2s_2l	Weight1 ^d	0.6437	0.0607	0.0204	0.0041	0.0467	0.6188	0.1006	0.0121	0.0075	0.6038	0.0856	0.0199	0.0105	0.6038	0.0856
	Weight2 ^e	0.6534	0.0604	0.0198	0.0040	0.0458	0.6192	0.1100	0.0115	0.0091		0.0946		0.0125		0.0946
	Weight3 ^f	0.6281	0.0606	0.0205	0.0041	0.0308	0.6192	0.0848	0.0136	0.0050		0.0694		0.0076		0.0694
2s_5l	Weight1 ^d	0.6929	0.0915	0.0343	0.0096	0.1150	0.6649	0.1702	0.0119	0.0144	0.6143	0.1195	0.0232	0.0295	0.6143	0.1195
	Weight2 ^e	0.6968	0.0958	0.0377	0.0106	0.1071	0.6655	0.1783	0.0115	0.0163		0.1271		0.0323		0.1271
	Weight3 ^f	0.6869	0.0836	0.0297	0.0079	0.1157	0.6654	0.1562	0.0131	0.0112		0.1051		0.0249		0.1051
5s_2l	Weight1 ^d	0.6286	0.0981	0.0316	0.0106	0.1564	0.6197	0.1642	0.0124	0.0116	0.5624	0.1069	0.0228	0.0275	0.5624	0.1069
	Weight2 ^e	0.6312	0.0991	0.0311	0.0108	0.1450	0.6192	0.1679	0.0119	0.0125		0.1111		0.0287		0.1111
	Weight3 ^f	0.6264	0.0930	0.0313	0.0096	0.1701	0.6216	0.1571	0.0137	0.0098		0.0978		0.0252		0.0978

^a Bias = the true AUC_T – the estimated AUC_T .

^b SD = Standard deviation.

^c Mean squared error = $\text{Bias}^2 + \text{SD}^2$.

^d $\omega_{k,k'} = \omega_{k,k} = 1$.

^e $\omega_{k,k'} = m_k + m_{k'}$ and $\omega_{k,k} = m_k + m_c$, where m_k is the sample size of subphenotype k and m_c is the sample size of controls.

^f $\omega_{k,k'} = \frac{\sqrt{m_k + m_{k'}}}{m_k \times m_{k'}}$ and $\omega_{k,k} = \frac{\sqrt{m_k + m_c}}{m_k \times m_c}$, where m_k is the sample size of subphenotype k and m_c is the sample size of controls.

g Probability of incorrectly combined subphenotypes.

$h_{2,2}$ corresponds to two disease models, each comprised a two-locus multiplicative interaction and an independent additive locus; 5_5 corresponds to two disease models, each comprised a two-locus multiplicative interaction, one two-locus threshold interaction, and an independent locus; both disease models of 2s_2l involved a two-locus multiplicative interaction and an independent locus, but the effect sizes are differed between two models; the first disease model of 2s_5l involved a two-locus multiplicative interaction and an independent locus, while the second disease model involved a two-locus multiplicative interaction, a two-locus threshold interaction, and an independent locus; the first disease model of 5s_2l involved a two-locus multiplicative interaction and an independent locus, while the second disease model involved a two-locus multiplicative interaction, a two-locus threshold interaction, and an independent locus.

Table 3

Mean, standard deviation, mean squared error of AUC_T in Simulation III

Models			AUC_T (MLR)				AUC_T (FROC-SUB)				AUC_T (FROC-ALL)			
Disease 1	Disease 2	Weight	Mean	Bias ^a	SD ^b	MSE ^c	Mean	Bias ^a	SD ^b	MSE ^c	Mean	Bias ^a	SD ^b	MSE ^c
1+2*3	4+5*6 ^g	Weight1 ^d	0.6770	0.0631	0.0212	0.0044	0.6770	0.0631	0.0212	0.0044	0.5700	0.1701	0.0260	0.0296
		Weight2 ^e	0.6739	0.0621	0.0210	0.0043	0.6739	0.0621	0.0210	0.0043		0.1659		0.0282
		Weight3 ^f	0.6815	0.0641	0.0217	0.0046	0.6815	0.0641	0.0217	0.0046		0.1755		0.0315
1+2*3	4+2*3 ^h	Weight1 ^d	0.6434	0.0510	0.0226	0.0031	0.6432	0.0511	0.0219	0.0031	0.6416	0.0527	0.0244	0.0034
		Weight2 ^e	0.6443	0.0509	0.0224	0.0031	0.6452	0.0501	0.0217	0.0030		0.0536		0.0035
		Weight3 ^f	0.6421	0.0510	0.0228	0.0031	0.6405	0.0526	0.0223	0.0033		0.0515		0.0033
1+2*3	1+2*3 ⁱ	Weight1 ^d	0.6681	0.0345	0.0234	0.0017	0.6151	0.0875	0.0163	0.0079	0.6682	0.0344	0.0233	0.0017
		Weight2 ^e	0.6680	0.0347	0.0235	0.0018	0.6217	0.0809	0.0172	0.0068		0.0344		0.0017
		Weight3 ^f	0.6681	0.0346	0.0235	0.0017	0.6056	0.0970	0.0152	0.0096		0.0344		0.0017

^a Bias = the true AUC_T – the estimated AUC_T .

^b SD = Standard deviation.

^c Mean squared error = $\text{Bias}^2 + \text{SD}^2$.

^d $\omega_{k,k'} = \omega_{k,k} = 1$.

^e $\omega_{k,k'} = m_k + m_{k'}$ and $\omega_{k,k} = m_k + m_c$, where m_k is the sample size of subphenotype k and m_c is the sample size of controls.

^f $\omega_{k,k'} = \sqrt{\frac{m_k + m_{k'}}{m_k \times m_{k'}}}$ and $\omega_{k,k} = \sqrt{\frac{m_k + m_c}{m_k \times m_c}}$, where m_k is the sample size of subphenotype k and m_c is the sample size of controls.

^g The two diseases were mutually independent.

^h The two diseases shared the same interactive loci.

ⁱ The two disease shared the same underlying disease model.

Table 4

Risk prediction models for T2D in *GEI-T2D*

Models	MLR Normal weight and overweight	FROC-SUB			MLR/FROC-SUB Obese	FROC-ALL
		Normal weight	Overweight			
Predictors	Alcohol consumption Smoking rs7901695 (TCF7L2) rs6931514 (CDKAL1)	Alcohol consumption Smoking rs4506565 (TCF7L2)	Alcohol consumption Smoking rs4506565 (TCF7L2) rs7756992(CDKAL1)		Alcohol consumption rs4607103 (ADAMTS9) rs564398(CDKN2A/2B) rs5215 (KCNJ11) rs4506565 (TCF7L2) rs8042680 (FTO)	Alcohol consumption Smoking
AUC (CI)	0.6149 (0.5941, 0.6324)	0.6041 (0.5774, 0.6309)	0.6281 (0. 6041, 0.6520)		0.6703 (0.6348, 0.7058)	0.5922 (0.5777, 0.6066)