Public Health Res Perspect 2011 2(2), 75–82
doi:10.1016/j.phrp.2011.07.005
pISSN 2210-9099 eISSN 2233-6052

ELSEVIER

ORIGINAL ARTICLE

Development of a Predictive Model for Type 2 Diabetes Mellitus Using Genetic and Clinical Data

Juyoung Lee [a,*], Bhumsuk Keam [b], Eun Jung Jang [c], Mi Sun Park [d], Ji Young Lee [a], Dan Bi Kim [d], Chang-Hoon Lee [e], Tak Kim [f], Bermseok Oh [g], Heon Jin Park [h], Kyu-Bum Kwack [i], Chaeshin Chu [c], Hyung-Lae Kim [j]

[a]*Division of Structural and Functional Genomics, Korea National Institute, Osong, Korea.*
[b]*Division of Biobank for Health Sciences, Korea National Institute, Osong, Korea.*
[c]*Division of Epidemic Intelligence Service, Korea Centers for Disease Control and Prevention, Osong, Korea.*
[d]*Division of Bio-Medical Informatics, Korea National Institute, Osong, Korea.*
[e]*Department of Internal Medicine and Lung Institute, Seoul National University College of Medicine, Seoul, Korea.*
[f]*Division of Quarantine Support, Korea Centers for Disease Control and Prevention, Osong, Korea.*
[g]*Department of Biomedical Engineering, School of Medicine, Kyung Hee University, Seoul, Korea.*
[h]*Department of Statistics, Inha University, Incheon, Korea.*
[i]*Medical Genomics Laboratory, Pochon CHA University, Seongnam, Korea.*
[j]*Center for Genome Research, Korea Centers for Disease Control and Prevention, Osong, Korea.*

Received: March 30, 2011
Revised: April 21, 2011
Accepted: May 1, 2011

KEYWORDS:
classification,
early predictive model,
single nucleotide polymorphism (SNP),
type 2 diabetes mellitus (T2DM)

Abstract
Objectives: Recent genetic association studies have provided convincing evidence that several novel loci and single nucleotide polymorphisms (SNPs) are associated with the risk of developing type 2 diabetes mellitus (T2DM). The aims of this study were: 1) to develop a predictive model of T2DM using genetic and clinical data; and 2) to compare misclassification rates of different models.
Methods: We selected 212 individuals with newly diagnosed T2DM and 472 controls aged in their 60s from the Korean Genome and Epidemiology Study. A total of 499 known SNPs from 87 T2DM-related genes were genotyped using germline DNA. SNPs were analyzed for significant association with T2DM using various classification algorithms including Quest (Quick, Unbiased, Efficient, Statistical tree), Support Vector Machine, C4.5, logistic regression, and K-nearest neighbor.
Results: We tested these models using the complete Korean Genome and Epidemiology Study cohort (n = 10,038) and computed the T2DM misclassification rates for each model. Average misclassification rates ranged at 28.2–52.7%. The misclassification rates for the logistic and machine-learning

*Corresponding author.
E-mail: jylee@cdc.go.kr

algorithms were lower than the statistical tree algorithms. Using 1-to-1 matched data, the misclassification rate of the statistical tree QUEST algorithm using body mass index and SNP variables was the lowest, but overall the logistic regression performed best.
Conclusions: The K-nearest neighbor method exhibited more robust results than other algorithms. For clinical and genetic data, our "multistage adjustment" model outperformed other models in yielding lower rates of misclassification. To improve the performance of these models, further studies using warranted, strategies to estimate better classifiers for the quantification of SNPs need to be developed.

1. Introduction

The prevalence of type 2 diabetes mellitus (T2DM) is increasing and the disease is becoming a worldwide epidemic [1]. T2DM is a complex disease affected by both genetic and environmental factors. Recent genetic association studies have provided convincing evidence that several novel loci and specific single nucleotide polymorphisms (SNPs) are associated with an increased risk of T2DM [2−6]. However, whether these SNPs have predictive value for future development of T2DM by an individual is controversial [7,8]. Hence development of an optimal and useful predictive model for T2DM using both genetic and clinical data is needed, because early prediction of T2DM can facilitate early control of blood glucose and lead to better clinical outcomes for patients with T2DM, or who may develop T2DM. However, there are few reports regarding prediction models that incorporate both genetic and clinical data.

We developed a new risk-prediction model for T2DM based on various statistical algorithms, and compared our model with preexisting models in terms of the error rate, also referred to as misclassification. We used SNPs and clinical data derived from a large community-based prospective cohort. The aims of this study were: (1) to develop a predictive model of T2DM using both genetic and clinical data; and (2) to compare the misclassification rates of our model versus other models.

2. Methods

2.1. Study population

The study samples were drawn from the Korean Genome and Epidemiology Study (KoGES), an ongoing prospective community-based epidemiological study in the communities of Ansung (rural) and Ansan (urban). From this study cohort ($n = 10{,}038$), we selected subjects who were aged in their 60s. From blood, it had measured fasting glucose, insulin, 1 hour and 2 hour glucose levels after the ingestion of 75 grams of glucose. Using these measures, we diagnosed new T2DM. We excluded 264 subjects who were already diagnosed with T2DM using questionnaire and selected newly diagnosed T2DM cases and age-matched controls from the KoGES cohort.

The diagnosis of T2DM was based on WHO criteria [9]. Control individuals had no past history of T2DM and had HbA1C values $<5.8\%$.

2.2. Selection of polymorphisms in candidate genes

Using the NCBI (National Center for Biotechnology Information) database and published reports, we selected 87 known candidate genes associated with T2DM. The selected genes included 15 genes involved in insulin metabolism, 8 genes involved in fatty acid binding/translocation, 13 genes involved in GLUT4 translocation, and 51 others. Subjects (control and T2DM) were genotyped by Taqman/Goldengate method, and 499 SNPs in 87 genes analyzed.

The association of a given SNP with T2DM were determined by χ^2 tests. We then selected significantly associated SNPs to optimize the predictive model and summarized the associations with allele types (dominant, recessive, heterozygous). Subsequently, we generated predictive models for T2DM using selected SNPs and clinical data. We then tested these models using data on the original study cohort ($n = 10{,}038$) and computed the misclassification rates for each model.

2.3. Statistical and classification analysis methods

We used a statistical decision tree classification algorithm to select SNPs because of the difficulty in optimizing high-risk SNPs.

We merged epidemiological, clinical, and genetic data to optimize classification. We then randomly divided the dataset into two parts, one for training or constructing the model and the other for testing the performance of the model (70% of the dataset was used for training and 30% for validation). The simulations resampled the data randomly and each simulation was repeated 100 times. We used various classification algorithms in this analysis based on the work of Ripley [10], Breiman [11], Hastie [12], and Quinlan [13] as follows. After first analyzing and screening for significantly associated SNPs using the χ^2 test, we used all the