Analysis of important factors in diabetes using machine learning algorithms and KoGES (Korean Genome and Epidemiology Study) data

Hark Soo SONG[1], Hynk KANG[1], Young Jin KIM[1], Ruda RHEE, KIM Sang Soo[2,3], KIM Jeong Mi[2,3], JANG Min Hee[2,3], YI Wook[2,3], RYANG Soree[2,3], KIM Minsoo[2,3], KIM In Joo[2,3], KIM Jinmi[4]

1) *Division of Basic Researches for Industrial Mathematics, Daejeon 34-047, KOREA*
2) *Division of Endocrinology and Metabolism, Department of Internal Medicine, Pusan National University Hospital, Pusan 49-241, KOREA*
3) *Biomedical Research Institute, Pusan National University Hospital, Pusan 49-241, KOREA*
4) *Division of Biostatistics, Clinical Trial Center, Biomedical Research Institute, Pusan National University Hospital, Pusan 49-241, KOREA*

Corresponding Author: Hark Soo SONG, hssong @gwmail.nims.re.kr

ABSTRACT

Recently, as the average life span of humans is prolonged due to the development of medical technology and the evolution of individuals' awareness of health, the prevalence of chronic diseases or complex diseases is rapidly increasing. In particular, in the case of chronic diseases such as diabetes and diseases that cause complications at the same time, the modern people's dietary habits and lifestyle patterns change, leading to a rapid upward curve. In this study, we aim to use machine learning methods to predict how the transition to diabetes will progress in the future if type 2 diabetes patients maintain their current diet and lifestyle. The data used in this study are large-scale medical examination records data, which is a Korean genome epidemiology study (KoGES, The Korean Genome and Epidemiology Study) conducted by the Korea Centers for Disease Control and Prevention. We had to improve the quality of the data in order to use machine learning methods. Improving the quality of the data is because it guarantees the accuracy of predictions and describes the reality well.
We will introduce methods and results for extracting variables closely related to diabetes and the process of data purification in order to extract statistical rules and useful information from large amounts of unrefined data through data mining methods.