SAT-LB121: Development of a Machine-Learning Method for Predicting New Onset of Diabetes Mellitus: A Retrospective Analysis of 509,153 Annual Specific Health Checkup Records

Akhiro Nomura. *Kanazawa University Graduate School of Medical Sciences*

Akihiro Nomura, MD, PhD[1], Sho Yamamoto, UG[2], Yuta Hayakawa, UG[2], Kouki Taniguchi, UG[2], Takuya Higashitani, MD[1], Daisuke Aono, MD[1], Mitsuhiro Kometani, MD, PhD[1], Takashi Yoneda, MD, PhD[1].

[1]Kanazawa University Graduate School of Medical Sciences, Kanazawa, Japan, [2]Kanazawa University, Kanazawa, Japan.

Diabetes mellitus (DM) is a chronic disorder, characterized by impaired glucose metabolism. It is linked to increased risks of several diseases such as atrial fibrillation, cancer, and cardiovascular diseases. Therefore, DM prevention is essential. However, the traditional regression-based DM-onset prediction methods are incapable of investigating future DM for generally healthy individuals without DM. Employing gradient-boosting decision trees, we developed a machine learning-based prediction model to identify the DM signatures, prior to the onset of DM. We employed the nationwide annual specific health checkup records, collected during the years 2008 to 2018, from Kanazawa city, Ishikawa, Japan. The data included the physical examinations, blood and urine tests, and participant questionnaires. Individuals without DM (at baseline), who underwent more than two annual health checkups during the said period, were included. The new cases of DM onset were recorded when the participants were diagnosed with DM in the annual check-ups. The dataset was divided into three subsets in a 6:2:2 ratio to constitute the training, tuning (internal validation), and testing datasets. Employing the testing dataset, the ability of our trained prediction model to calculate the area under the curve (AUC), precision, recall, F1 score, and overall accuracy was evaluated. Using a 1,000-iteration bootstrap method, every performance test resulted in a two-sided 95% confidence interval (CI). We included 509,153 annual health checkup records of 139,225 participants. Among them, 65,505 participants without DM were included, which constituted36,303 participants in the training dataset and 13,101 participants in each of the tuning and testing datasets. We identified a total of 4,696 new DM-onset patients (7.2%) in the study period. Our trained model predicted the future incidence of DM with the AUC, precision, recall, F1 score, and overall accuracy of 0.71 (0.69-0.72 with 95% CI), 75.3% (71.6-78.8), 42.2% (39.3-45.2), 54.1% (51.2-56.7), and 94.9% (94.5-95.2), respectively. In conclusion, the machine learning-based prediction model satisfactorily identified the DM onset prior to the actual incidence.