Abstract

Diabetes mellitus is a metabolic disorder characterized by chronic hyperglycemia, which may cause serious harm to many of the body's systems. Diabetes is a deadly pandemic which presents a significant burden on healthcare systems worldwide, and will continue to do so as its global prevalence rises rapidly (particularly type 2 diabetes). In developed countries, the rising prevalence is primarily driven by population aging, lifestyle changes and greater longevity of diabetes patients. Diabetes can be managed effectively when detected early.
Unfortunately, early detection proves difficult as the time between onset and clinical diagnosis may span several years. Furthermore, estimates indicate that over one third of diabetes patients in developed countries are undiagnosed.

We investigated the potential of Belgian health expenditure data as a basis to build a cost-effective population-wide screening approach for (type 2) diabetes mellitus, aspiring to improve secondary prevention by speeding up the diagnosis of patients in order to initiate treatment before the disease has caused irrevocable damage. We used health expenditure data collected by the National Alliance of Christian Mutualities – the largest social health insurer in Belgium. This data comprises basic biographic information and records of all refunded medical interventions and drug purchases, thus providing a long-term longitudinal overview of over 4 million individuals' medical expenditure histories.

Screening was formulated as a binary classification task, in which diabetes patients represent the positive class. Due to the nature of the problem and limitations of health expenditure data, we were unable to identify a set of known negatives (patients without diabetes). Hence, we had to learn classifiers from positive and unlabeled data. During this project we made two contributions to this subdomain of semi-supervised learning: (i) a novel learning method which is robust to false positives and (ii) an approach to evaluate classifiers using traditional metrics without known negatives in the test set. Additionally, we mapped the survival of patients starting various antidiabetic pharmacotherapies and developed two open-source machine learning packages: one for ensemble learning and another to automate hyperparameter search.

We built a screening method with competitive performance to existing state-of-the-art approaches. This exceeded our expectations, since health expenditure data omits most info about the typical risk factors used by other screening methods (BMI, lifestyle, genetic predisposition, . . .). As such, the combination of health expenditure data and additional information about risk factors is a promising avenue for future research in screening for diabetes mellitus. Finally, our approach has a very low operational cost since we only used readily-available data, which effectively removes one of the key barriers of population-wide screening for diabetes.