健診ビッグデータによる多疾患発症予測モデル構築

内野 詠一郎
京都大学大学院　医学研究科
人間健康科学系専攻　医療情報 AI システム学講座

専門分野・キーワード：腎臓内科学、機械学習
自己紹介：電子カルテデータや健診データなどの健康医療リアルワールドデータ×機械学習 による知識発見やスマートな医療の実現を目指しています。

健康長寿社会の実現に向け、疾患発症前の段階における各種疾患の発症予測と、それによる予防介入が期待されている。近年、機械学習技術を大規模データに適用することによる予測モデルの構築が着目されており、これまで様々な疾患に対する予測モデル構築やそれによる関連因子の同定が報告されている。一方、生活習慣病をはじめとする多様な疾患を一括して取り扱うことは従来のコホートデータでは難しく、個人に対する包括的な疾患発症予測やそれによる予防介入は十分実現されていない。

本研究では、「岩木健康増進プロジェクト」による健康診断データを使用し、各種疾患の 3 年以内の新規発症を予測するモデルを作成した。2005 年から 2017 年の 13 年分のデータから、社会背景、生活習慣、各種計測値等の 2804 項目およびその時系列での差分項目の計 5319 項目を使用し、糖尿病、高血圧症、慢性腎臓病等の計 20 疾患の新規発症を予測対象とした。モデルの構築には Random Forest、Light GBM、XGBoost を使用した。

モデルの構築に使用しなかった 20%のテストデータセットによる評価において、AUROC 値で概ね 0.7～0.9 程度の性能が見られた（図 1）。本データセットおよび機械学習技術による疾患発症予測が可能であることが示されたとともに、今後、本モデルにおける変数寄与度等の解析によって、各種疾患の新規の関連因子や最適な介入点の発見につながる可能性が示唆された。

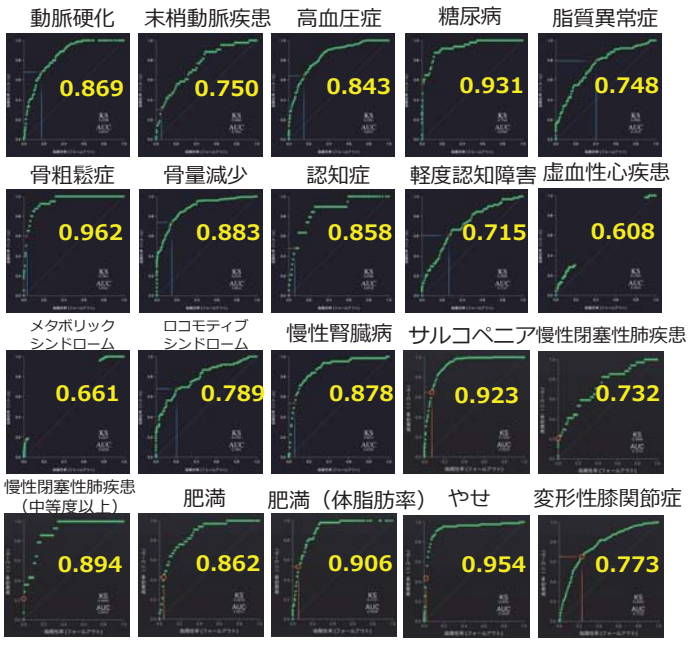

図 1. 20 疾患の予測モデルの ROC 曲線と AUC 値