健診・医療・介護ビッグデータを用いた疾病発症の予測

○桑野将司　中井健太郎　岩田千加良　南野友香　森山卓（鳥取大学）

Predictions of Disease Onsets using Medical Big Data

* M. Kuwano, K. Nakai, C. Iwata, Y. Minamino, and T. Moriyama (Tottori University)

概要ー　本研究では, 鳥取県国民健康保険団体連合会が保有する国保データベースシステムに蓄積されている健診･医療･介護ビッグデータを用いて疾病発症の予測および発症要因の特定を目的とする. 糖尿病，高血圧症，脂質異常症，および 3 つの疾病をまとめて定義した生活習慣病の 4 種類を分析対象に，被保険者の将来 3 年間における疾病発症確率をディープニューラルネットワークにより予測し，PFI 指標と決定木分析によって疾病発症要因の特定を行った.

キーワード:　疾病発症予測，深層学習，Permutation Feature Importance，決定木分析

1　研究の背景と目的

医療分野においては，従来の仮説駆動型医科学に加え，ビッグデータとして蓄積された情報に AI 技術を活用して新たな知見を得るデータ駆動型医科学による研究の推進が期待されている．本研究では，健診・医療・介護ビッグデータを用いて疾病発症の予測および発症要因の特定を目的とする．予測対象疾病は「糖尿病」，「高血圧症」，「脂質異常症」に加えて，3 つの疾病をまとめて定義した生活習慣病の 4 種類とし，（i) ディープニューラルネットワーク（以降は, DNN）による疾病発症の予測，（ii）PFI（Permutation Feature Importance）による説明変数の重要度分析，（iii）決定木分析による要因特定の 3 段階の分析を行う.

2　使用データの概要

鳥取県国民健康保険団体連合会が保有する国保データベースシステムに蓄積されている健診・医療・介護ビッグデータを分析対象とする. 2017 年 1 月 1 日から 2020 年 12 月 31 日までの 48 ヶ月間のうち，1）2017 年以前に対象疾病発症していないこと, 2) サンプル内に欠損値が存在しない完全データであることの 2 つの条件を設定し，サンプルを抽出する．また，データセットの作成において，質問票データに欠損値が多数含まれることから，サンプル数と説明変数の数の違いが予測精度に及ぼす影響を把握するために，（1）質問票データを含むデータセットと（2）質問票データを含まないデータセットの 2 パターンについて分析する.

3　予測モデルの構築

被保険者ごとの将来 3 年間における疾病発症確率を算出する DNN モデルを構築する．モデルの過学習の抑制を目的として k-分割交差検証，不均衡データへの対処を目的としてオーバーサンプリングを採用した.

構築した DNN モデルの有効性を再現率および適合率の観点から検証した結果，生活習慣病の再現率は 55.1%, 適合率は 60.3%と, 高い予測精度を得た. また，疾病別の予測精度は，生活習慣病が最も高く，次いで糖尿病，脂質異常症，高血圧症となった．さらに，生活習慣病の疾病予測には，質問票データを含むデータセットを使用した方が予測精度は高い結果となった.

4　PFI による疾病発症要因の分析

DNN は高い予測精度を有する一方で, 入力層から出力層までの過程がブラックボックスであることから，予測結果に対する解釈が困難であるというデメリットが存在する．そこで，損失関数の増加量に着目した手法である PFI を用いて，予測精度に対する各説明変数の重要度を測った. Fig. 1 にそれぞれ生活習慣病の PFI 平均値の上位 5 項目を示す．糖尿病の検査項目の 1 つである HbA1c が最も予測精度に影響を与えていることが明らかになった．しかし，増加量の比較基準となる本来の損失関数に比べて，これら要因の寄与度が小さいことから，必ずしも予測精度に対してある 1 つの要因のみの影響が大きいとはいえない．すなわち，疾病の発症において，複数の説明変数が相互に作用していると考えられ，個別の説明変数のみに着目した予測結果の解釈は困難であることが示唆された.

5　決定木分析による疾病発症要因

説明変数の相互作用を考慮した疾病発症要因を明らかにするために，決定木分析を適用する．なお，目的変数は「DNN から得られた分類である真陰性（TN），偽陽性（FP），偽陰性（FN），真陽性（TP）」の 4 項目を設定した．生活習慣病に関する分析の結果，対象疾病に関わる検査値の HbA1c, 収縮期血圧，LDL コレステロールが保健指導判定値以上で, 57kg 以上の被保険者に対しては, DNN で正確に疾病発症を予測できることが明らかとなった．また，PFI 同様，HbA1c が有意な変数として検出されたことから，HbA1c が生活習慣病の発症予測に対して影響を及ぼしていることを裏付けるとともに，他の説明変数と相互に作用していることが示された.

6　結論

本研究では，健診・医療・介護データを用いて疾病発症予測および分類要因を特定した．今後は，欠損値が多い質問票データへの対処方法を検討することで，予測精度を向上する必要がある.

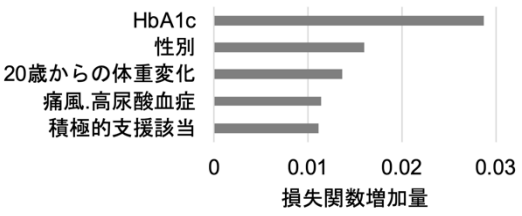

Fig. 1: 生活習慣病の PFI 平均値の上位 5 項目