氏　　名	大岡 忠生
博士の専攻分野の名称	博　士（医　学）
学位記番号	医工博4甲　第277号
学位授与年月日	令和元年6月12日
学位授与の要件	学位規則第4条第1項該当
専　　攻　　名	生体制御学専攻
学位論文題名	Artificial Intelligence Approach to Type 2 Diabetes Risk Prediction and Exploration of Predictive Factors: Applying Machine Learning Technique to Large-scale Health Checkup Data （人工知能技術を用いた2型糖尿病のリスク予測と疾患予測因子の探索：大規模健康診断データへの機械学習技術の適用）
論文審査委員	委員長　教　授　秋山 真治 委　員　准教授　柏木 賢治 委　員　准教授　古屋 文彦

学位論文内容の要旨

（研究の目的）

近年、機械学習法を中心とした人工知能技術が医療の分野に導入され始め、様々な成果や研究結果を残している。しかし、ディープラーニングを代表とする人工知能技術を用いて予測した結果の多くは、その予測の理由を説明できない事（「ブラックボックス問題」と呼ばれる）や、大量のデータが予測に必要とされる事が、医学研究や医療現場に人工知能技術を取り入れる上での大きな障壁となっている。人工知能技術を代表する手法の中で、予測の理由を解釈しやすく、比較的少ないデータでの解析を可能とする機械学習手法であるランダムフォレストという解析手法が存在する。ランダムフォレストでは線形モデルを使わない事から、多重共線性の問題が起きにくい事が知られており、時系列変化を含む多変数を予測モデルに投入する事が可能である。ランダムフォレストが既存のリスクモデル研究で用いられる既存手法よりも高い精度で疾病を予測する事が出来るのであれば、理想的な疾患のリスク予測や疾患予測因子の同定を行う事が可能となる。そこで、本研究では機械学習法ランダムフォレストを大規模健康診断結果に適用する事で糖尿病予測モデルを作成し、既存手法との予測能の比較や予測因子の妥当性について検討を行う事で、疫学データを用いた医学研究における機械学習法（ランダムフォレスト）の適切な適用方法やそのメリット、注意点を模索する事を目的とした。

（方法）

1999年4月から2009年3月までに山梨厚生連健康管理センターで健康診断を受けた、延べ168,206人の受診者を研究対象とした。連続二年間の結果を用いて次年の糖尿病リスクを予測する目的のも

と、三年間連続で健康診断を受けた受診者のうち、糖尿病の治療中でなく、HbA1cが6.5%未満である42,908人を解析の対象とした。予測モデルの作成にあたっては、HbA1cが1年後に0%以上、0.2%以上、0.4%以上、0.6%以上、0.8%以上、1.0%以上上昇したか否かによる二値変数を目的変数として設定し、単一年度の健康診断検査値と前年度からの変化値合計97変数を説明変数の候補として、各解析手法において6つの予測モデルを作成した。本研究ではランダムフォレスト（RF）、多変量ロジスティック回帰分析（MLR）、限定的ランダムフォレスト（LRF）、限定的多変量ロジスティック回帰分析（LMLR）の4つの解析手法を用いて、各目的変数を取った時のROC(Receiver Operating Characteristic)曲線を描画し、AUC(Area Under the Curve)を算出する事で各モデルの予測能の比較を行った。また、各予測モデルにおける重要な予測因子を比較する為、ランダムフォレスト(RF)において変数重要度を、多変量ロジスティック回帰分析(MLR)において標準偏回帰係数を算出し、それぞれ比較を行った。

（結果）

４つの解析手法におけるROC曲線の比較では、6種いずれの予測モデルにおいてもランダムフォレスト（RF）が最も高い予測力を示した。また、ROC曲線から算出したAUCも全予測モデルにおいてランダムフォレスト(RF)が有意に高い数値を示した。ランダムフォレスト(RF)を用いた予測モデルにおける変数重要度は、（前年度からの）HbA1c変化、HbA1c、血糖値、血糖値変化、体重、ALP、血小板数、CRP変化、ALP変化、中性脂肪の順に高かった。一方、多変量ロジスティック回帰分析（MLR）における標準偏回帰係数を算出すると、MCH、MCV、MCHC、ヘマトクリット、ヘモグロビン、総コレステロール、LDLコレステロール、血糖値、HbA1c、HbA1c変化が重要な変数として挙げられた。

（考察）

ランダムフォレストを用いたモデルで高い予測精度が出た理由として、ランダムフォレストが分類の過程において線形モデルを使用しない為、多重共線性の問題を起こさず多くの変数をバランスよく考慮することが出来た事が考えられる。また、単一年度の健康診断結果のみでなく前年度からの変化を考慮して予測を行う事で、より高精度に糖尿病リスクを予測できた事は、リスクモデルを作成する際に継時的な変化を考慮すべき可能性を示唆している。そのため継時的な変数を考慮した上で、多重共線性を起こさずに高精度の予測能を示しているランダムフォレストモデルは、既存手法より正確に予測因子を同定できている可能性がある。実際、本研究でランダムフォレストモデルにより示された重要な予測因子の殆どは、既存研究により糖尿病リスクとの関連が示唆されている。本研究では、十分な数の対象者に対して機械学習法を用いることで高精度のリスクモデルを作成し、妥当性の高い予測因子を同定することが出来たが、研究対象集団の選び方や糖尿病治療が自己申告である事、数理研究上示唆されるランダムフォレスト特有の問題点を克服出来ていない事は、本研究の限界である。今後は他の医療データへの適用を通じて、機械学習法を用いた最適な解析方法や正しい解釈法を検討していく事が望まれる。

（結論）

機械学習法ランダムフォレストを大規模健康診断データに適用することで、既存手法よりも高い精度で糖尿病リスク予測モデルを作成する事が出来た。また、予測因子として妥当な変数が選ばれていた。適切に機械学習法を用いることで、既存手法よりも高精度なリスク予測や疾患予測因子の同定が可能となる事が本研究により示唆された。今後は他の医療データへの適応を通して、技術を

用いる際の利点と注意点について更なる検証が望まれる。

論文審査結果の要旨

１．学位申請論文の学術的意義

近年、医療分野でも人工知能は目覚ましい研究成果を上げており、今後、分子生物学の分野から公衆衛生の分野まで、医学研究の広い範囲への適用の可能性が見込まれるが、本研究は、その流れに沿って、人工知能の一手法である機械学習技術を公衆衛生分野に応用する妥当性と可能性を検討した研究である。

静的データ（動的データ）を機械的学習によって分類（予測）する技術であるランダムフォレストは、多変数かつ欠損値の多いデータに適しており、かつ、症例数が少ない医療データのような偏ったデータに対しても分類（予測）の妥当性が期待されている手法である。本研究では、大規模健康診断データに対して糖尿病の発症リスク因子HbA1cの１年後の変化を予測するランダムフォレストモデルを作成し、既存の予測モデル（多変量ロジスティック回帰分析や単年度データのみを用いた予測）と精度比較を行い、予測においてランダムフォレストが優位であること示した。また、ランダムフォレストモデルの各因子の変数重要度を算出する事で、糖尿病の予測に関わる因子を新しい形で提示できることを示した。

２．学位申請論文の新しい点

回帰を使わずに予測因子の同定を行う手法であるランダムフォレストが、①回帰分析よりも高い精度で疾患予測をする事ができ、②臨床的に妥当性の高い予測因子を与えることを明らかにし、ランダムフォレストモデルによって理想的なリスクモデルの作成や疾患予測因子が探索できることを示した。

３．実験（数値統計解析）の信頼性

口頭試問で疑問点を質問したりする中で、申請者は人工知能の医療分野への応用に通暁しているだけでなく、本論文で使ったランダムフォレストモデルや人工知能の主流であるディープラーニングなども自らプログラムを作成して実行しており、本論文の数値統計解析手順も詳細にチェックしていることが判明した。したがって、本申請論文において導かれた結論は十分に信頼できる。

４．学位申請論文の改善点など

ランダムフォレストモデルと既存の予測モデルを大規模健康診断データに対してそれぞれ適用した結果を比較・検討した研究論文であることをより明確に記し、ランダムフォレストモデルやその他の人工知能関連技術による疾患予測に関しては今後の研究に期待する。