様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ５ 月 １２ 日現在

機関番号：１３５０１
研究種目：若手研究
研究期間：2019～2021
課題番号：１９Ｋ１９４３３

研究課題名（和文）疫学データに対する人工知能技術適用の枠組みの検証と提案

研究課題名（英文）Validation and proposal of a framework for the application of artificial intelligence techniques to epidemiological data

研究代表者
大岡　忠生（Ooka, Tadao）

山梨大学・大学院総合研究部・助教

研究者番号：４０８０３９８７

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：健康診断施設における約20年間の健康診断結果のデータを活用する事で、2型糖尿病における重要な指標であるHbA1cが急上昇する受診者を、前年度の健康診断結果から高精度に予測する人工知能モデルの開発に成功した。また、これらのモデルを検証する事で、2型糖尿病の発症を予測するうえで重要となる血糖値以外の要素（例：コレステロール値、血圧）を同定した。更に、このモデルを発展させて、過去の健康診断結果から1年後・3年後の健康診断結果を高精度に予測する人工知能モデルも併せて開発した。今後は同モデルを実際の健康診断や保健指導の現場で活用する事で受診者の健康が促進するかを確認するランダム化比較試験を実施していく。

研究成果の学術的意義や社会的意義
様々な機械学習モデルを疫学データに活用する事で、疫学データへの人工知能（機械学習）技術適応の枠組みの検証を行うことが出来た。また、将来の健康診断結果を高精度に予測する機械学習モデルの開発にも成功した。今後は、開発した予測モデルをどのように使うか、研究の枠組みをどのように活用していくかを検討するために、ランダム化比較試験を含めた更なる検討を進めていく。

研究成果の概要（英文）：By utilizing data from about 20 years of health checkups at health checkup facilities, we have succeeded in developing an artificial intelligence model that can accurately predict who will have a sharp rise in HbA1c, an important indicator of type 2 diabetes, based on the results of the previous year's health checkups. By validating these models, we identified factors (e.g., cholesterol levels, blood pressure) that are important in predicting the onset of type 2 diabetes.
Furthermore, by developing this model, we have also developed an artificial intelligence model that can accurately predict the results of health checkups one and three years from now, based on the results of past health checkups.
In the future, a randomized controlled trial will be conducted to confirm whether the model can be used in actual health checkups and health guidance to promote the health of examinees.

研究分野：先制医療

キーワード：疾患予測モデル　糖尿病　健康診断　機械学習　ランダムフォレスト

様　式　C－１９、F－１９－１、Z－１９（共通）

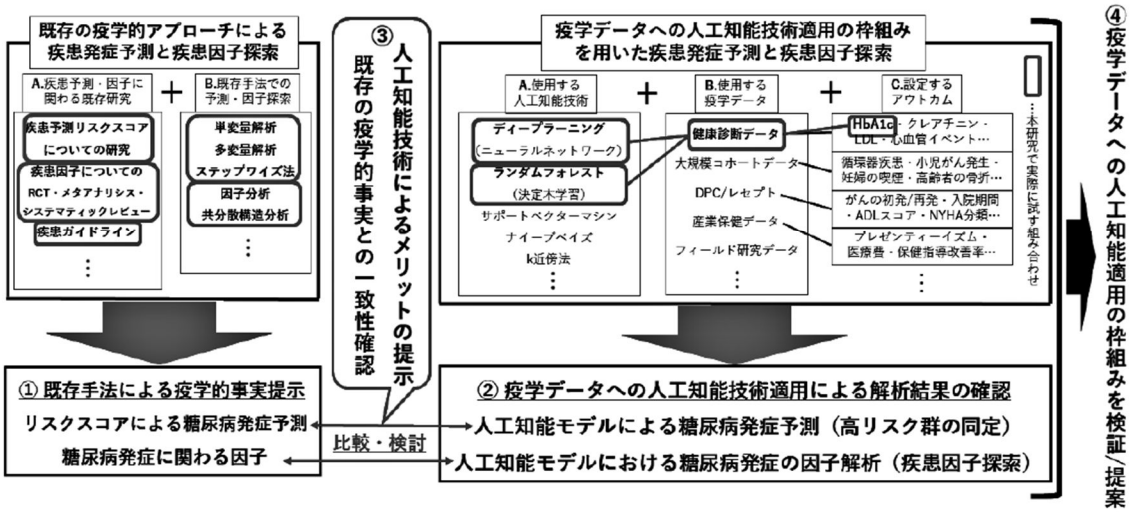
１．研究開始当初の背景

医療現場や疫学研究において収集される情報量は年々増加している。予防医療を推進していく上で、健康診断や遺伝子情報を代表とする医療ビッグデータや大規模疫学データを用いて早期段階で正しい介入を行う事が、今後の医療体制において肝要である。近年、画像診断を中心とした人工知能技術が医療現場に導入されようとしているが、予防医療や疫学研究において人工知能技術を積極的に活用している研究は少ない。その理由として、①人工知能技術が疫学研究のどのような場面で有効であるか不明である②人工知能技術の疫学データへの適用に妥当性があるか不明である③研究者が新たに人工知能技術を習得する難しさ、等が挙げられる。さらに疫学データの解析を適切に行い、結果を正しく解釈するには、医学的知識や疫学の知識や経験が求められ、統計の専門家が率先して解析を行う事が難しい分野でもある。人工知能技術を疫学研究に導入する際には、人工知能技術を含むデータ解析と医学や疫学の知識を組み合わせた研究が求められており、専門外の研究者でも人工知能技術を疫学データに適用できる枠組みの整備が求められる。

２．研究の目的

本研究では、疫学データへの人工知能技術適用による疾患発症予測と疾患因子探索を実際に行い、既存の疫学的事実と比較を行うことで人工知能技術による解析に妥当性があるのか、既存手法に対してメリットが存在するかを確認し、疫学データへの人工知能技術適応の枠組みの検証と提案を行うことを目的とする。具体的には、HbA1c(長期血糖値)により判定した糖尿病発症の有無をアウトカムとし、まず既存の疫学的アプローチにより糖尿病の発症予測と疾患因子に関わる研究を収集し、多変量解析や因子分析といった既存疫学手法による探索を行う。次に、人工知能技術による解析を実施し、先に提示した疫学的事実と人工知能技術による解析結果との一致性や人工知能技術適応による解析のメリットを確認する。以上の試行を通し、人工知能技術適用の際の注意点や利点を踏まえて、疫学データへの人工知能技術適応の枠組みの検証と提案を行う。まとめると、図１における①②③の検討を通して、④で示す疫学データへの人工知能技術適応の枠組みを検証し、提案する事を最終的な目的とする。

図1　本研究の全体概要図



３．研究の方法

本研究は、予め用意された大規模健康診断データに人工知能技術を適用する事で糖尿病の発症予測と予測因子の同定を行い、既存の疫学的アプローチにより解明されていた事実との一致性を確認した上で、人工知能技術による追加の解析メリットを提示することで、疫学データに人工知能技術を導入する際の枠組みを検証し、提案するものである。既存の疫学的アプローチとしては、糖尿病を予測するリスクスコアについての研究や糖尿病発症の予測因子についてのエビデンスレベルの高い研究（RCT,メタアナリシス等）を参照するとともに、本研究で用いる健康診断データに多変量解析とステップワイズ法を適応する事で糖尿病の発症予測や標準偏回帰係数の比較による因子同定、共分散構造分析による因子構造の確認を行う。適応する人工知能技術としてはRandom ForestとDeep Learningの二手法を採用し、必要に応じて他手法も活用する。誰もが使用できる研究の枠組みの提案を前提に、無料の統計解析ソフトR(最新版)を解析の中心に

用いる。統計数理研究所と各研究協力者の技術支援を受けて各人工知能手法の導入や調節を行い、糖尿病発症予測に最適化したモデルを作成する。完成したモデルを用いて、使用した変数の重要度を比較し、先に提示した疫学的事実との比較を行う。以上の解析実施と解析結果の検証を通して、疫学データへの人工知能適応の枠組みを検討、提案する。

４．研究成果

（１） 健康診断結果から HbA1c の急上昇を予測する機械学習モデルの開発

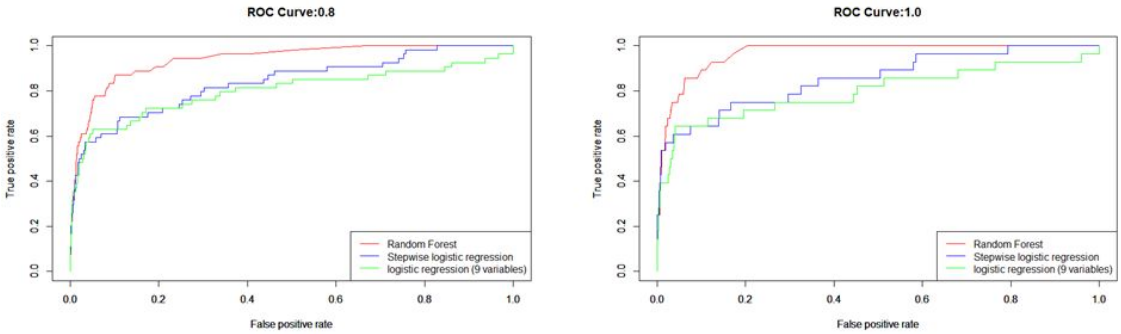
1999 年 4 月から 2009 年 3 月までに山梨県厚生連健康管理センターで健康診断を受診した、延べ 168,206 人の受診者を対象者とし、各受診者の健康診断における全 215 項目の検査結果から、次年度の HbA1c の急上昇の有無を予測する機械学習モデルを開発し、比較した。目的変数には、前年度から次の年度にかけての HbA1c の変化量を異なる閾値（model1:0, model2:+0.2, model3:+0.4, model4:+0.6, model5:+0.8, model6:+1.0）で二値としたものを使用し、目的変数ごとに 6 つのモデルを作成した。説明変数には健康診断で検査あるいは問診された 215 変数のうち、受診者全員が測定する一般的な項目 51 変数を使用した。その内、46 変数については前年度からの変化量も使用した。予測モデルはランダムフォレスト(RF model)、ロジスティック回帰(MLR model)を作成し、ROC 曲線と AUC による比較を行った。

結果として、ランダムフォレストを用いたモデルは、ロジスティック回帰モデルよりも全ての目的変数において予測精度が高かった。（図２・３）特に前年度から HbA1c が 0.8 以上あるいは 1.0 以上上昇する対象者を予測するモデルにおいては、高い精度を実現した。

図２　各モデルの精度

		Best model on ROC curve				Best model on ROC curve	
	AUC	Sensitivity	Specificity		AUC	Sensitivity	Specificity
RF model1	0.719	0.714	0.617	MLR model1	0.699	0.648	0.648
RF model2	0.716	0.608	0.720	MLR model2	0.711	0.648	0.668
RF model3	0.743	0.607	0.778	MLR model3	0.734	0.629	0.729
RF model4	0.864	0.804	0.823	MLR model4	0.817	0.748	0.773
RF model5	0.940	0.870	0.898	MLR model5	0.840	0.685	0.889
RF model6	0.967	0.929	0.877	MLR model6	0.854	0.750	0.834

図３　HbA1c急上昇予測モデルの精度比較



以上の成果につき、2021 年度に査読付き国際雑誌にて出版を行った。[1]

（２） 将来の健康診断結果を予測するモデルの開発

上記と同様のデータを用いて、16 種類の健康診断項目＜体重、腹囲、BMI、血圧(収縮期,拡張期)、肝機能(AST,ALT,γ-GTP)、脂質(中性脂肪,LDL/HDL コレステロール)、血糖値(空腹時血糖,HbA1c)、腎機能(クレアチニン,eGFR)、尿酸値＞を高精度に予測するモデルの作成を行い、比較を行った。具体的には、2 年連続の健康診断結果から 1 年後、3 年後の健康診断結果を予測する機械学習モデル（ロジスティック回帰、ラッソ回帰、ランダムフォレスト、XGBoost、Deep Neural Network）を作成し、比較を行った。

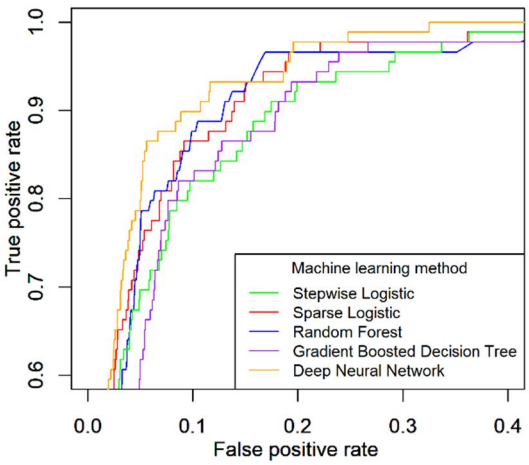
図４　将来の健康診断予測モデルの精度比較

Method	ROC-AUC	Optimal point		PR-AUC	Optimal point		
		Sensitivity	Specificity		F-score	Precision	Recall
Stepwise Logistic Regression	0.939	0.910	0.825	0.236	0.333	0.258	0.472
Sparse Logistic Regression	0.955	0.933	0.850	0.270	0.363	0.355	0.371
Random Forest	0.949	0.966	0.831	0.210	0.304	0.227	0.461
Gradient Boosted Decision Tree	0.931	0.933	0.806	0.151	0.249	0.168	0.483
Deep Neural Network	0.964	0.933	0.884	0.322	0.405	0.333	0.517

結果としては、Deep Neural Network を用いたモデルが最も高い ROC-AUC と PR-AUC となり、次いでラッソ回帰（Sparse Logistic Regression）が高い ROC-AUC と PR-AUC となった。ロジスティック回帰、ラッソ回帰、ランダムフォレストを用いて、予測に対して重要度の高い変数を検討したところ、HbA1c と空腹時血糖に加え、LDL コレステロール、HDL コレステロール、拡張期・収縮期血圧、喫煙などが重要な変数として提示されていた。

同結果は、2021 年度の国際疫学会（InternationalEpidemiological Association）にて発表され、同学会の学会誌に掲載された。[2]

図５　各予測モデルの精度比較



（３）健康診断予測モデルの活用による受診者の健康行動促進の有無の検討

上記にて開発した 16 項目の健診項目を予測するモデルを実際に健康診断施設の受診者に適用し、予測結果を基に保健指導を実施する事で、実際に受診者の健康増進に繋がるのかどうかを検討するランダム化比較試験を計画し、2021 年度から実際に開始されている。

＜引用文献＞

1. Ooka T, Johno H, Nakamoto K, Yoda Y, Yokomichi H, Yamagata Z. Random Forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan. BMJ Nutr Prev Health. 2021;4(1):140-148.
2. Ooka T, Yokomichi H, Yamagata Z. Intelligence Approaches to Type 2 Diabetes Risk Prediction and Exploration of Predictive Factors, International Journal of Epidemiology, Volume50, Issue Supplement_1, September 2021, dyab168.515

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Tadao Ooka, Hisashi Johno, Kazunori Nakamoto, Yoshioki Yoda, Hiroshi Yokomichi, Zentaro Yamagata	４．巻 -
２．論文標題 Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan	５．発行年 2021年
３．雑誌名 BMJ Nutrition, Prevention & Health	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Ooka Tadao、Yokomichi Hiroshi、Yamagata Zentaro	４．巻 50
２．論文標題 425Artificial Intelligence Approaches to Type 2 Diabetes Risk Prediction and Exploration of Predictive Factors	５．発行年 2021年
３．雑誌名 International Journal of Epidemiology	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/ije/dyab168.515	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Ooka Tadao、Yokomichi Hiroshi、Yamagata Zentaro
２．発表標題 Artificial Intelligence Approaches to Type 2 Diabetes Risk Prediction and Exploration of Predictive Factors
３．学会等名 The World Congress of Epidemiology 2021（国際学会）
４．発表年 2021年

１．発表者名 大岡忠生、横道洋司、山縣然太朗
２．発表標題 機械学習を活用した将来の健康診断検査値の予測方法の検討
３．学会等名 第31回日本疫学会学術総会
４．発表年 2021年

１．発表者名
大岡忠生、横道洋司、山縣然太朗

２．発表標題
Deep Learning を活用して健康診断結果から糖尿病発症を予測する方法の検討

３．学会等名
第79回日本公衆衛生学会総会

４．発表年
2020年

１．発表者名
大岡忠生、横道洋司、山縣然太朗

２．発表標題
人工知能技術を活用した2型糖尿病のリスク予測手法の検証と疾患予測因子の探索

３．学会等名
第78回日本公衆衛生学会総会

４．発表年
2019年

１．発表者名
大岡忠生、横道洋司、山縣然太朗

２．発表標題
機械学習技術を用いて健康診断結果から糖尿病発症を予測する方法の検討

３．学会等名
第30回日本疫学会学術総会

４．発表年
2020年

１．発表者名
大岡忠生、日野英逸、横道洋司、山縣然太朗

２．発表標題
予防医療分野における疫学データへの機械学習技術活用について　～スパースモデリングを活用した糖尿病発症予測と予測因子探索～

３．学会等名
統計数理研究所共同利用研究集会 ~統計的機械学習の新展開~

４．発表年
2020年

〔図書〕　計1件

１．著者名 大岡忠生	４．発行年 2021年
２．出版社 医学書院	５．総ページ数 6ページ(pp. 115-120)
３．書名 公衆衛生	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関