

令和 2 年 6 月 16 日現在

機関番号：32643

研究種目：挑戦的研究（萌芽）

研究期間：2017～2019

課題番号：17K19930

研究課題名（和文）学習・検証・実践の3フェーズに基づく高血圧・脳卒中発症予測のための人工知能構築

研究課題名（英文）Construction of artificial intelligence to predict incidence of hypertension and stroke based on machine learning, verification, and practice phases.

研究代表者

大久保 孝義（Ohkubo, Takayoshi）

帝京大学・医学部・教授

研究者番号：60344652

交付決定額（研究期間全体）：（直接経費） 4,900,000円

研究成果の概要（和文）：株式会社JMDCが保有する健診データを用い、高血圧発症者と非発症者が同数になるようアンダーサンプリングを行った対象者で機械学習を行い、発症予測のための人工知能を構築した。次に大迫コホート研究のデータを用い、この予測能を測った。結果、ニューラルネットワークよりもロジスティック回帰分析の人工知能の方がややバランス良く高血圧発症と非発症を分類できていた。しかし、いずれの分類能も不十分と考えられた。そこで変数のカテゴリ化等を実施し精度向上を図ったが、大きな改善は認められなかった。脳卒中発症予測についても同様に検討したが、JMDCで構築した人工知能の大迫研究対象者におけるF値は極めて低値であった。

研究成果の学術的意義や社会的意義

JMDCデータで構築した高血圧・脳卒中発症予測の人工知能を大迫研究データに適用することは困難と考えられた。これは学習と検証に用いたデータに含まれる対象者特性の相違が原因と考えられる。傾向スコアマッチングによる両データの特性を一致させる、データのスケール変換などにより大迫研究データと互換性が取れるJMDCデータの再構築をする、といった前処理に関する今後の検討の必要性を明らかにした点で、本研究は萌芽研究として一定の意義を有するものと考えられる。

研究成果の概要（英文）：For the prediction of hypertension incidence within the next 5 years, the artificial intelligence (AI) was developed based on annual health check-up database from JMDC Inc. We then selected the same number of participants with and without hypertension incidence by the under-sampling method, respectively. We assessed the predictive value of the AI by applying it to data from the Ohasama cohort study. Although the AI developed by the logistic regression method showed better predictive value for incident hypertension than that developed by the neural network method, adequate predictive value was not observed from these two AIs. Categorization of variables, addition of other variables, or adjusting the parameter of neural network model did not significantly enhance the predictive value of AI. We also developed the AI for the prediction of stroke. However, the stroke prediction model from the JMDC database revealed low F value when it was applied to the data from the Ohasama cohort study.

研究分野：疫学・予防医学

キーワード：高血圧 機械学習 成人保健

1. 研究開始当初の背景

高血圧は、日本で現在 4,300 万名が有する疾患である。死因の上位を占める脳心血管疾患の最大の原因であり、高血圧、またそれに続く脳卒中発症予測システムに基づく早期予防対策実施は急務である。

ニューラルネットワークに基づく人工知能が、古典的な線形回帰モデルよりも強く高血圧発症を予測したとの海外の報告がある (ROC 曲線下面積: 0.90 vs 0.73) [Huang S, et al. Hypertens Res. 2010]。これは独自の栄養調査結果を人工知能とした結果であり、日本人に適用できない結果であるが、日本の一般的な健診項目を用いることで高血圧・脳卒中発症を予測する人工知能を構築できる可能性がある。

2. 研究の目的

本研究の目的は、

- ・ 機械学習を行い、高血圧および脳卒中の 5 年間発症予測のための人工知能を構築すること
- ・ 長期コホート研究のデータを用い、人工知能の妥当性を検証、そして実践で用いることである。

3. 研究の方法

【研究データとデザイン】

機械学習用のデータとして、まず、株式会社 JMDC が保有する健康保険組合被保険者の健診データを用いた。本研究では、株式会社 JMDC より受領している最低 5 回分 (年 1 回) 健診を受診した 388,973 名のデータを使用した。そのうち、初回受診時から 5 年以内に高血圧または脳卒中を新規に発症した対象を抽出した。JMDC データを用いた機械学習で人工知能の構築が可能な場合、佐久研究の 2008-2013 年に複数回人間ドックを受診した対象の空腹時採血・面接式問診で得られた高精度データを用いることを検討した。

妥当性の検証には、岩手県花巻市大迫町の住民を対象とした長期前向きコホート研究である大迫研究のデータを用いた。大迫研究では、大迫町 4 地区を 1 年 1 地区調査し、4 年サイクルで家庭血圧、身長、既往歴・現病歴や治療歴などの基礎情報、血液検査データ、および各種動脈硬化検査データを捉え続けている。岩手県脳卒中登録事業のデータを基に正確な脳卒中発症状況を捉えている。4 年後の家庭高血圧発症および 5 年以内の脳卒中発症を捉え、前述の人工知能を適用させた。

【調査項目】

JMDC 健診データの健診情報から、年齢、body mass index (BMI)、現在喫煙、現在飲酒、脳血管疾患既往、心臓病既往、および降圧薬服用情報を抽出した。問診情報と血液検査データを用い、脂質代謝異常症を LDL コレステロール 140 mg/dL 、HDL コレステロール $< 40 \text{ mg/dL}$ 、中性脂肪 150 mg/dL 、または問診で脂質代謝異常症治療薬服用有りと回答されていた場合と定義した。同様に、糖尿病を空腹時血糖 126 mg/dL 、随時血糖 200 mg/dL 、または問診で糖尿病治療薬有りと回答されていた場合と定義した。高血圧発症を、5 年後の健診で測定された随時収縮期/拡張期血圧が $140/90 \text{ mmHg}$ 以上または降圧治療有りと定義した。脳卒中発症を 5 年後の問診での脳卒中既往歴の有無で判断した。

大迫研究では、これまで収集された家庭血圧データと共に、本研究期間内に追加収集されたデータを統合し、家庭血圧の時系列データを構築した。家庭血圧値は、朝 1 機会 1 回目の測定値を採用した。各研究期間で 4 週間の測定を住民へ依頼し、その期間に測定された測定値を基に、家庭血圧値を個人内平均値、日間変動を個人内測定値の標準偏差を平均値で除した変動係数 (coefficient of variation [CV], %) として算出した。その他、JMDC データと同じ臨床情報を抽出し、各種疾患についても同じく定義した。

【解析方法】

初回健診時のベースラインデータを入力層、アウトカムである 5 年後の高血圧・脳卒中発症の有無を出力層とした機械学習を実施した。学習器として、ニューラルネットワーク (MLPClassifier) を使用したが、比較対照として線形モデルであるロジスティック回帰分析も同時に使用した。ニューラルネットワークは学習モデルに入れた因子のスケールの影響を強く受けるため、連続変量については平均 0・分散 1 となるよう標準化した。パラメーターサーチとして、10 回の交差検証に基づくグリッドサーチを実施した。

JMDC データで機械学習および交差検証により学習器の妥当性を検証し、その後、JMDC データで機械学習した人工知能を大迫研究データに適用させて有用性を検討した。評価指標として、ア

ウトカムおよび非アウトカムに対する、適合率、再現率、およびF値を算出した。なお、再現率は対象アウトカムの陽性反応適中度、再現率は感度と同じ算出方法であり、F値は再現率と適合率の調和平均である。要約統計量について、連続変数を平均値±標準偏差で示した。

4. 研究成果

データの確認

JMDC データと大迫研究データの解析用データセットを構築し、対象者特性の違いを検討するため、両データにおける加齢と血圧レベルの関連を検討した。JMDC データにおける随時収縮期血圧は、加齢とともにほぼ直線的に上昇し、70歳時点の男性および女性では129.3 mmHg および女性では129.1 mmHgであった。大迫研究データでもほぼ同様の血圧推移が認められたが、70歳の男性および女性の家庭収縮期血圧レベルはそれぞれ136.9 mmHg および137.7 mmHgであった。家庭血圧が随時血圧よりも平均で5 mmHg高いことを考慮しても、大迫研究データに比べ、JMDC データには血圧レベルが低い集団が含まれていると考えられた。これは、JMDC データに含まれる対象が、健康保険組合に加入者（大企業に勤める成人およびその扶養家族）であり、比較的健康的な集団が含まれているためと考えられた。大迫研究において、血圧及び各種検査指標と脳卒中発症や臓器障害との関連を明らかにし、血圧指標およびアウトカムの精度が高いことを示した。

高血圧・脳卒中発症予測発症予測に関する検討

JMDC データより、データ欠損が無く、脳心血管疾患既往のない非高血圧者216,199名のデータを抽出した。まず、このJMDC データの対象者で後述のものと同様の機械学習を実施し、高血圧・脳卒中発症予測の人工知能を作成したが、大迫研究への汎化性は極めて低かった。これはJMDC データと大迫研究の平均年齢に大きな差があること（45歳 vs 68歳）が影響していると考えられた。

そこで、JMDC データの対象を60歳以上の6,826名に限った。この対象における男性の割合は60.0%、平均年齢±標準偏差は63.1±2.8歳、収縮期/拡張期血圧は119.0±11.9/72.5±8.8 mmHgであり、高血圧発症者は22.4%（1,529名）であった。機械学習に用いるため、高血圧発症者と非発症者が同数になるようアンダーサンプリングを行い、最終的に3,058名を抽出した。モデルに投入した因子として、男性、年齢、BMI、喫煙、飲酒、糖尿病、脂質異常症、収縮期血圧、および拡張期血圧を用いた。パラメータサーチにより、最適化手法としてlbfsgs、隠れ層は2層でニューロン数は10:100ユニット、L2正則化のペナルティを示すalpha=0.001が選択された。これらパラメータを利用し、10回交差検証を実施した（表1）。

表1. 交差検証による高血圧発症の機械学習予測結果（ニューラルネットワーク）

評価指標	評価指標の平均（最小-最大）	
	非発症(0)に対する評価	発症(1)に対する評価
適合率	0.64 (0.60-0.69)	0.68 (0.62-0.75)
再現率	0.71 (0.63-0.82)	0.61 (0.56-0.66)
F値	0.68 (0.62-0.73)	0.64 (0.59-0.69)

対象を60歳以上に限ったJMDC データを用いて交差検証10回を実施した。

次に、モデルに投入した因子のスケールへの依存が少なく汎用性の高いロジスティック回帰分析による機械学習を実施した。グリッドサーチの結果に基づき、正則化の強さを指定するパラメータCを1.0とし、交差検証を実施したところ表2の結果が得られた。

表2. 交差検証による高血圧発症の機械学習予測結果（ロジスティック回帰）

評価指標	評価指標の平均（最小-最大）	
	非発症(0)に対する評価	発症(1)に対する評価
適合率	0.60 (0.54-0.65)	0.62 (0.55-0.68)
再現率	0.65 (0.54-0.72)	0.57 (0.49-0.63)
F値	0.62 (0.56-0.68)	0.59 (0.52-0.64)

対象を60歳以上に限ったJMDC データを用いて交差検証10回を実施した。

最終的に、アンダーサンプリング後のJMDC対象者3,058名で機械学習を行った。これらの人工知能を用い、直近8年間の間に大迫研究に2期連続（4年間隔で2回）受診した非高血圧者85名（年齢：65.8歳、男性19名、高血圧新規発症者32名）を対象に、上記の人工知能の予測能を測った（表3）。結果、ニューラルネットワークよりもロジスティック回帰分析の人工知能の方が、ややバランスの良い高血圧発症と非発症を分類できていた。しかし、これら評価指標の値から、分類能は十分ではないと考えられた。

次に、変数のカテゴリ化、因子の追加（LDL・HDLコレステロールや肝機能指標 -GTP・GPT・

GOT など) そしてニューラルネットワークについては隠れ層やニューロン数等のパラメータの調整を実施し、分類能の向上を図ったが、高血圧発症・非発症の予測能に大きな改善は認められなかった。大迫コホートデータについて、機械学習に用いた随時血圧を、5 mmHg 加算した家庭血圧値に入れ替えた検討もしたが、予測能に改善は認められなかった。

表 3. 大迫研究データにおける高血圧発症の予測結果

評価指標	高血圧発症予測能に関する評価指標の値			
	ニューラルネットワーク		ロジスティック回帰	
	非発症(0)	発症(1)	非発症(0)	発症(1)
適合率	0.68	0.39	0.76	0.45
再現率	0.23	0.84	0.42	0.78
F 値	0.34	0.54	0.54	0.57

JMDC データで機械学習したモデルを大迫研究データに当てはめた結果である。

脳卒中発症予測の人工知能の構築も検討した。脳卒中既往情報より、JMDC データには 5 年以内の脳卒中新規発症者 142 名(平均 63.5 歳)が確認された。同数の非発症者のデータを加え、計 284 名のデータで高血圧発症予測の人工知能と同じフローで機械学習を実施し、脳卒中発症予測の人工知能を構築した。

大迫研究データからは、データ欠損や脳心血管疾患既往歴のない 718 名のうち、18 名が 5 年以内に脳卒中を発症していた(発症者の平均年齢 70.5 歳)。この対象者に JMDC で構築した脳卒中発症予測の人工知能を適用したが、発症に対する F 値は 0.1 未満と極めて低値であった。

以上の検討から、JMDC データで構築した高血圧・脳卒中発症予測の人工知能を大迫研究データに適用することは困難と考えられた。これは学習と検証に用いたデータに含まれる対象者特性の相違が原因と考えられる。傾向スコアマッチングによる両データの特性を一致させる、データのスケール変換などにより大迫研究データと互換性が取れる JMDC データの再構築をする、といった前処理に関する検討が必要と考えられる。

機械学習の学習器について、高血圧・脳卒中発症予測能の精度はニューラルネットワークよりもロジスティック回帰分析モデルの方がやや良好であったが、大幅な改善は認められなかった。その他の学習器である、決定木、ランダムフォレストや Ridge・Lasso 回帰分析、またはこれらを組み合わせたアンサンブル学習も検討し、より汎化性の高い人工知能の構築についても検討が必要と考えられた。

5. 主な発表論文等

〔雑誌論文〕 計6件（うち査読付論文 6件/うち国際共著 0件/うちオープンアクセス 2件）

1. 著者名 Sato M, Asayama K, Kikuya M, Inoue R, Tsubota-Utsugi M, Obara T, Murakami K, Matsuda A, Murakami T, Nomura K, Metoki H, Imai Y, Ohkubo T.	4. 巻 40(1)
2. 論文標題 Nocturnal blood pressure decline based on different time intervals and long-term cardiovascular risk: the Ohasama Study.	5. 発行年 2018年
3. 雑誌名 Clin Exp Hypertens.	6. 最初と最後の頁 1-7
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/10641963.2016.1259324.	査読の有無 有
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 -
1. 著者名 Sato M, Murakami T, Asayama K, Hirose T, Kikuya M, Inoue R, Tsubota-Utsugi M, Murakami K, Matsuda A, Hara A, Obara T, Kawasaki R, Nomura K, Metoki H, Node K, Imai Y, Ohkubo T.	4. 巻 82(8)
2. 論文標題 N-Terminal Pro-B-Type Natriuretic Peptide Is Not a Significant Predictor of Stroke Incidence After 5 Years- The Ohasama Study.	5. 発行年 2018年
3. 雑誌名 Circ J.	6. 最初と最後の頁 2055-2062
掲載論文のDOI（デジタルオブジェクト識別子） 10.1253/circj.CJ-17-1227.	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 Sato M, Asayama K, Murakami T, Kikuya M, Metoki H, Imai Y, Ohkubo T.	4. 巻 42(1)
2. 論文標題 Stroke risk due to partial white-coat or masked hypertension based on the ACC/AHA guideline's blood pressure threshold: the Ohasama study.	5. 発行年 2019年
3. 雑誌名 Hypertens Res.	6. 最初と最後の頁 120-122
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41440-018-0133-2.	査読の有無 有
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 -
1. 著者名 佐藤倫広, 村上尚, 小原拓, 辰巳友佳子, 高畠恭介, 原梓, 浅山敬, 今井潤, 菊谷昌浩, 大久保孝義, 目時弘仁.	4. 巻 54(3)
2. 論文標題 大規模健診時血圧データに基づく加齢に伴う血圧推移に関する縦断解析.	5. 発行年 2019年
3. 雑誌名 日本循環器病予防学会誌.	6. 最初と最後の頁 163-169
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 オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 Sato M, Metoki H, Asayama K, Murakami T, Inoue R, Tsubota-Utsugi M, Matsuda A, Hirose T, Hara A, Obara T, Kikuya M, Nomura K, Hozawa A, Imai Y, Ohkubo T.	4. 巻 8(15)
2. 論文標題 Age-Related Trends in Home Blood Pressure, Home Pulse Rate, and Day-to-Day Blood Pressure and Pulse Rate Variability Based on Longitudinal Cohort Data: The Ohasama Study.	5. 発行年 2019年
3. 雑誌名 J Am Heart Assoc.	6. 最初と最後の頁 e012121
掲載論文のDOI (デジタルオブジェクト識別子) 10.1161/JAHA.119.012121.	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 佐藤倫広, 村上任尚, 小原拓, 辰巳友佳子, 高畠恭介, 原梓, 浅山敬, 今井潤, 菊谷昌浩, 大久保孝義, 目時弘仁.	4. 巻 54
2. 論文標題 大規模健診時血圧データに基づく加齢に伴う血圧推移に関する縦断解析.	5. 発行年 2019年
3. 雑誌名 日本循環器病予防学会誌	6. 最初と最後の頁 163-169
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13件(うち招待講演 0件/うち国際学会 2件)

1. 発表者名 佐藤倫広, 村上任尚, 浅山敬, 菊谷昌浩, 井上隆輔, 坪田恵, 小原拓, 村上慶子, 松田彩子, 原梓, 野村恭子, 目時弘仁, 今井潤, 大久保孝義.
2. 発表標題 一般地域住民における家庭血圧に基づく高血圧の分布—大迫研究—.
3. 学会等名 第66回東北公衆衛生学会.
4. 発表年 2017年

1. 発表者名 佐藤倫広, 村上任尚, 浅山敬, 菊谷昌浩, 坪田恵, 原梓, 松田彩子, 井上隆輔, 野村恭子, 目時弘仁, 今井潤, 大久保孝義.
2. 発表標題 米国高血圧新基準と一般地域住民における白衣・仮面高血圧の分布: 大迫研究.
3. 学会等名 第67回東北公衆衛生学会.
4. 発表年 2018年

1. 発表者名 平塚貴子, 村上任尚, 佐藤倫広, 小宮山貴将, 齋藤翔, 大井孝, 遠藤耕生, 浅山敬, 菊谷昌浩, 井上隆輔, 坪田恵, 村上慶子, 目時弘仁, 今井潤, 服部佳功, 大久保孝義.
2. 発表標題 歯数と家庭血圧日間変動との関連-大迫研究一.
3. 学会等名 第29回血圧管理研究会.
4. 発表年 2017年

1. 発表者名 川崎良, 浅山敬, 佐藤倫広, 菊谷昌浩, 井上隆輔, 坪田恵, 松田彩子, 村上任尚, 原梓, 野村恭子, 目時弘仁, 竇澤篤, 今井潤, 大久保孝義.
2. 発表標題 血圧と網膜血管径の関連および高血圧治療薬の影響 大迫研究
3. 学会等名 第29回日本疫学会学術総会.
4. 発表年 2019年

1. 発表者名 須賀明子, 布恵子, 佐藤倫広, 浅山敬, 小林信一郎, 世古俊明, 山中珠美, 松本知沙, 菊谷昌浩, 野村恭子, 井上茂, 今井潤, 大久保孝義.
2. 発表標題 眼底所見と認知機能の関連: 大迫研究.
3. 学会等名 第29回日本疫学会学術総会.
4. 発表年 2019年

1. 発表者名 奈良井大輝, 佐藤倫広, 村上任尚, 菊地ひかり, 中山晋吾, 高畠恭介, 廣瀬卓男, 佐藤和奏, 浅山敬, 菊谷昌浩, 野村恭子, 竇澤篤, 目時弘仁, 今井潤, 大久保孝義.
2. 発表標題 白衣高血圧と左室肥大の関連における家庭血圧レベルの影響: 大迫研究.
3. 学会等名 日本高血圧学会 第8回臨床高血圧フォーラム.
4. 発表年 2019年

1. 発表者名 Bursztyn M, Kikuya M, Asayama K, Satoh M, Gavish B, Ohkubo T.
2. 発表標題 ELASTIC COMPONENT OF 24-HOUR OF PULSE PRESSURES PREDICTS CARDIOMUSCULAR AND TOTAL MORTALITY IN THE OHASAMA STUDY.
3. 学会等名 29th European Meeting on Hypertension and Cardiovascular. (国際学会)
4. 発表年 2019年

1. 発表者名 佐藤倫広, 廣瀬卓男, 中山晋吾, 村上任尚, 高畠恭介, 大久保孝義, 森建文, 目時弘仁.
2. 発表標題 降圧薬服用有無別の血圧レベルと慢性腎臓病発症リスクとの関連: 大規模健診データに基づく検討.
3. 学会等名 第42回日本高血圧学会総会.
4. 発表年 2019年

1. 発表者名 村上任尚, 佐藤倫広, 小宮山貴将, 大井孝, 遠藤耕生, 平塚貴子, 浅山敬, 菊谷昌浩, 井上隆輔, 坪田恵, 目時弘仁, 寶澤篤, 今井潤, 服部佳功, 大久保孝義.
2. 発表標題 重度歯周病罹患歯の保有数と血圧との関連 - 大迫研究 -
3. 学会等名 第31回血圧管理研究会
4. 発表年 2019年

1. 発表者名 Satoh M.
2. 発表標題 AGE-RELATED TRENDS IN BLOOD PRESSURE BASED ON LARGE-SCALE HEALTH CHECKUP DATA USING LONGITUDINAL ANALYSIS.
3. 学会等名 29th European Meeting on Hypertension and Cardiovascular. (国際学会)
4. 発表年 2019年

1. 発表者名 佐藤 倫広, 村上 任尚, 小原 拓, 辰巳 友佳子, 高畠 恭介, 原 梓, 浅山 敬, 今井 潤, 大久保 孝義, 目時 弘仁.
2. 発表標題 Defined Daily Doseと比較した日本の降圧薬処方用量 レセプトデータに基づく検討
3. 学会等名 第21回日本医薬品情報学会総会・学術大会
4. 発表年 2018年

1. 発表者名 佐藤 倫広, 村上 任尚, 小原 拓, 辰巳 友佳子, 高畠 恭介, 原 梓, 浅山 敬, 今井 潤, 大久保 孝義, 目時 弘仁.
2. 発表標題 大規模健診時血圧データに基づく加齢に伴う健診時血圧の推移に関する検討
3. 学会等名 第54回日本循環器病予防学会
4. 発表年 2018年

1. 発表者名 佐藤 倫広, 村上 任尚, 小原 拓, 辰巳 友佳子, 原 梓, 浅山 敬, 今井 潤, 大久保 孝義, 目時 弘仁.
2. 発表標題 大規模健診時血圧データに基づくリアルワールドにおける降圧治療前後の血圧状況に関する前向き検討
3. 学会等名 第7回 臨床高血圧フォーラム
4. 発表年 2018年

〔図書〕 計0件

〔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
研究 分担者	佐藤 倫広 (Satoh Michihiro) (70717892)	東北医科薬科大学・医学部・助教 (31305)	