

リアルワールドデータを用いた腎機能低下予測モデルの開発と 機械学習によるモデル改善の検討

東北医科薬科大学医学部 衛生学・公衆衛生学教室
講師 佐藤 倫広

(共同研究者)

帝京大学医学部 衛生学公衆衛生学講座	主任教授	大久保 孝義
帝京大学医学部 衛生学公衆衛生学講座	講師	辰巳 友佳子
東北医科薬科大学医学部 衛生学・公衆衛生学教室	教授	目時 弘仁
東北医科薬科大学内科学第三（腎臓内分泌内科）教室	教授	森 建文
東北医科薬科大学内科学第三（腎臓内分泌内科）教室	助教	中山 晋吾

はじめに

高齢化に伴い腎機能の低下が進行すると、末期腎不全、さらには循環器疾患につながる。我が国の透析患者数は1970年の9,491名から2019年には344,640名にまで増加しており、その前段階にある腎機能低下の早期予防は急務の課題である。軽度の腎機能低下は自覚症状がないため、早期の発見が重要である。

近年では健康情報の電子化とその蓄積が進み、リアルワールドデータと呼ばれる非臨床試験で得られる情報の利活用が進められている。この大規模データを利用することで、個々人の将来の疾患リスク予測が可能となる。そこで本研究の目的は、リアルワールドデータとして日本人の健診・レセプトデータおよび一般住民の長期腎機能推移データを用い、腎機能低下予測モデルの開発、および、この開発における機械学習の有用性を検討することとした。

将来の腎機能低下予測モデルが開発され、健診等に導入されれば、より早期に具体的な値をもって高リスク対象者への指導ができる。ひいては、より効果的な生活習慣の是正および腎機能管理につながる。予測モデル構築には、異なる予測モデル作成と妥当性検証の2つのデータが必要である。本研究では、長期の家庭血圧データも並行して収集する一般住民データも有しており、特性の異なるコホートをを用いた予測モデル妥当性の検証が可能である。

結 果

本研究が利用した大規模リアルワールドデータとして、DeSCヘルスケア株式会社が提供するデータベース（以下、DeSCデータベース）を使用した。DeSCデータベースには、①大企業とされる民間企業の被雇用者とその扶養家族が加入する組合管掌健康保険、②自営業者・退職者が加入している国民健康保険、③75歳以上の高齢者が加入している後期高齢者医療

制度の健診・レセプトデータが含まれている。

DeSCデータベースに含まれる30歳以上で推算糸球体濾過量（estimated Glomerular Filtration Rate: eGFR）15 mL/min/1.73 m²以上の糖尿病治療薬服用中患者を対象とし、患者背景、各種糖尿病治療薬、降圧薬の使用の有無を用いた傾向スコアマッチングを行ったうえで、ジペプチジルペプチダーゼ4（Dipeptidyl Peptidase-4: DPP-4）阻害薬既存服薬群と非服薬群対象者を抽出した。eGFR 45 mL/min/1.73 m²以上の患者65,375名での検討の結果、ベースラインからのeGFRの低下度は、DPP-4阻害薬既服用者は非服用者と比較して2年目および3年目で有意に緩徐であった。以上から、糖尿病治療中患者において、血糖降下薬の服用薬剤の違いが腎機能に大きく影響を及ぼすことが明らかとなった⁽¹⁾。

次に、ベースラインのBody Mass Index（BMI）とeGFR低下との関連を明らかにした。女性ではeGFR傾斜の低下とBMIが直線的に関連していた一方、男性では調整なしでもU字型の関連が観察された。これらの結果は、共変量で調整した後も変わらなかった（図1）。

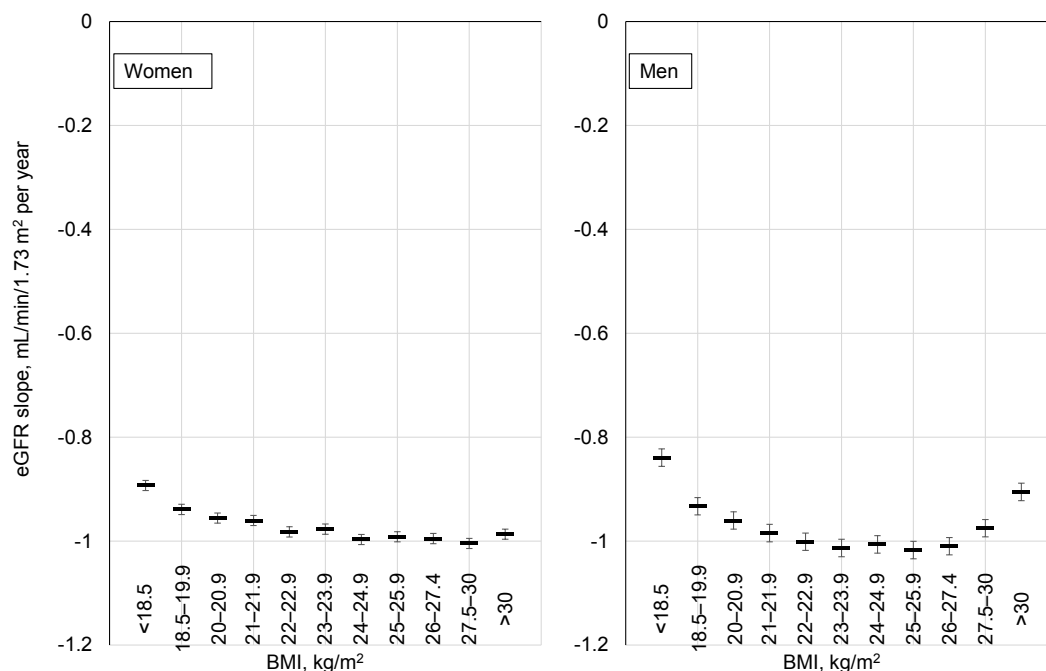


図1 ベースラインBMI別のeGFR年間変化度

eGFR年間変化度は、ベースライン時の年齢、性、喫煙歴、飲酒歴、収縮期血圧、HbA1c、LDLコレステロール、降圧薬の使用、糖尿病治療薬の使用、脂質低下薬の使用、脳卒中および虚血性心疾患既往、およびeGFRで調整されている。

次に、血圧分類とeGFR低下度の年間推移の関連を線形混合モデルで捉えたところ、血圧が高い群ほど大きくeGFRが低下していた（図2）。その他、尿酸治療薬と腎機能への影響も検討したが、糖尿病の有無で薬剤による腎機能への差異は一部認められたものの、全体での解析では尿酸治療薬による腎機能への影響は限定的と考えられた。

糖尿病治療薬⁽¹⁾および降圧薬^(2, 3)の腎機能への影響を考慮し、eGFR年間低下度予測のための機械学習と人工知能作成では、糖尿病治療薬および降圧薬の使用が追跡中に無く、蛋白尿も無いeGFRが45～120 mL/min/1.73 m²の313,974名を対象として、eGFR年間低下度を予測するための機械学習モデルの探索を行った。機械学習に入れる説明変数として、ベースライン時の年齢、性、BMI、喫煙、飲酒、収縮期血圧、HbA1c、LDLコレステロール、脂質降下薬の使用有無、脳卒中既往、冠動脈疾患既往、およびeGFRを使用した。先のBMIとeGFR年間低下度とのU型関連を考慮するため、BMIを<18.5、25～30、および≥30 kg/m²にカテゴリ化したダミー変数をモデルに投入し、18.5～24.9 kg/m²群を基準とした計算とした。さらに、性以外の変数については、性との交互作用項を作成し、同時にモデルに投入した。

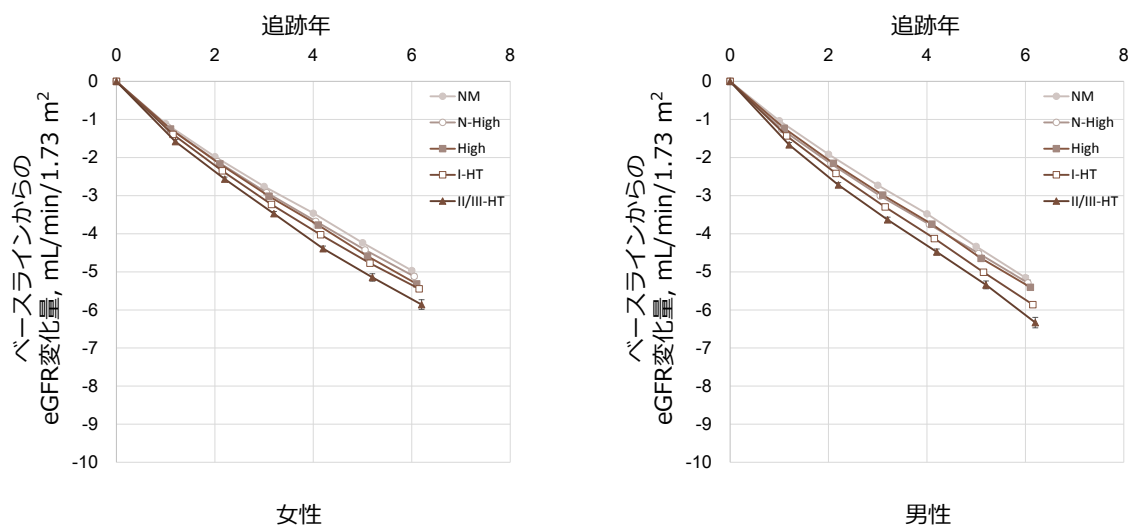


図2 ベースライン血圧分類別のeGFR年間変化度

eGFR年間変化度は、ベースライン時の年齢、BMI、喫煙、飲酒、HbA1c、LDL、降圧薬使用、糖尿病治療薬使用、脂質異常症治療薬使用、脳卒中既往、冠動脈疾患既往、データ内健診受診回数、eGFR、保険種別、および受診年度で調整されている。血圧分類は日本高血圧学会が定める高血圧治療ガイドライン2019に準拠している⁽⁴⁾。NM=正常血圧、N-High=正常高値血圧、High=高値血圧、およびHT=高血圧。

北海道・東北地域を検証コホートとして除外し、それ以外の地域のデータでeGFR年間低下度（目的変数）予測のための機械学習コホートとして解析を行った。説明変数を標準正規分布にする前処理を行い、集団を5分割の設定にした交差検証（クロスバリデーション）を行った。機械学習モデルとして、線形回帰、リッジ回帰、ラッソ回帰、勾配ブースティング決定木(XGBoost)、およびニューラルネットワークを利用した。5分割交差検証の結果、最も二乗平均平方根誤差（RMSE）が小さく、決定係数が大きいXGBoostモデルがeGFR年間低下度の予測能が良いと判断された。

表1 機械学習モデルのモデル適合度

モデル	平均RMSE	決定係数 (R-squared)
線形回帰 (lm)	0.9977798	0.02418418
リッジ回帰 (ridge)	0.9977868	0.02380693
ラッソ回帰 (lasso)	0.9977829	0.02388755
勾配ブースティング決定木 (XGBoost)	0.9906723	0.04200168
ニューラルネットワーク (nnet)	1.3370760	1.888577e-05

目的変数をeGFR年間低下度としたモデルの結果を示す。RMSE: Root Mean Squared Error

最終的に、機械学習コホート全体でXGBoostモデルに基づくeGFR年間低下度予測のための人工知能を構築し、検証コホートである北海道・東北地域のデータに適用したところ、予測能の決定係数は0.042であった。これは、このモデルがeGFR年間低下度の分散の4.2%のみを説明できることを意味している。XGBoostモデルにおいて特徴量重要度が最も高かった上位3つの変数は、ベースライン時のeGFR、年齢、および収縮期血圧であった。全特徴量の重要度の合計を100%とした場合、これらの変数の相対的重要度はそれぞれ40.8%、32.7%、および4.5%であった。

岩手県花巻市大迫町で実施される大迫コホート研究においても前述の機械学習による腎機能低下予測モデルを適用したが、eGFR年間低下度の予測能は限定的であった。この原因の一つとして、先の解析で重要度が比較的高かった血圧の測定方法の差異を考え、家庭血圧と健診時血圧に基づく血圧分類と長期腎機能推移を線形混合モデルで検討した。尿蛋白陰性の岩手県花巻市大迫町住民1,766名(平均61.3歳、男性34.9%)において、家庭血圧は腎機能低下と明瞭に関連した一方、健診時血圧に基づく血圧分類と腎機能低下の関連は不明瞭であった。

考 察

大規模リアルワールドデータを利用し、機械学習を用いて腎機能予測モデルの構築を検討した。その結果、XGBoostが最も予測能が高いモデルであった。XGBoostは、決定木モデルを基本とし、勾配ブースティングの原理に基づいて複数の弱学習器を作成し、それらの予測を組み合わせることで高精度な予測を実現するアルゴリズムである。そのために単純な線形モデルよりも複雑な予測が可能だった可能性がある。また、正則化アルゴリズムも組み込まれており、過学習も抑えられた可能性がある。一方で、より複雑なニューラルネットワークモデルのeGFR年間低下度の予測能が低かった理由は、過学習であった可能性が高い。ただし、データの前処理の見直しやファインチューニング等を行えばニューラルネットワークでも有効な予測モデルを構築できる可能性があるため、今後のさらなる検討が望まれる。

複雑な機械学習により、eGFR年間低下度の予測能は高まったものの、XGBoostであってもeGFR年間低下度の決定係数は4%に過ぎなかった。この決定係数の値は一般的に低く、

eGFR年間低下度の予測が困難であることが示唆されている。この一因は、eGFR測定と各リスク因子の偶然誤差の大きさであった可能性がある。本研究の解析対象者において腎機能低下と最も寄与が高かった因子は、ベースライン時eGFR、年齢、そして血圧レベルであった。ベースライン時eGFRについては、高いほどその後のeGFR低下度が大きいことを捉えている。今回使用した説明変数のうち、ベースライン時eGFRの相対的重要度が約40%と非常に高かったことは、平均への回帰の効果が極めて強く影響していたことを示唆している。

ベースライン時eGFR以外で、最もeGFR年間低下度の予測に寄与していた修正可能なりスク因子は血圧であった。しかし、健診で測定される健診時血圧は、防御・警鐘反応に由来する白衣効果が含まれ、再現性と信頼性が低い。加えて、日本で広く実施される特定健診や定期健診では、日本高血圧学会が推奨する測定方法とは全く異なる方法で測定されており、例えば常時2回測定している実施団体は調査参加施設の18.8%に過ぎなかったとも報告されている⁽⁵⁾。このようなデータベースでは捉えられない血圧測定の誤差が、予測モデル構築の障害となった可能性がある。この血圧測定の問題は、本研究で利用したデータベースに限ったことではなく、非試験条件下で実施された健診データを利用したデータベース全てに共通する⁽⁶⁾。したがって、健診現場でのより精度の高い血圧やその他健診項目の測定が必要である。

大迫研究を利用し、家庭血圧と健診時血圧の両方の腎機能低下予測能を比較したところ、より精度や予後予測能が高いとされる家庭血圧の方が良好であった。この結果も、機械学習で予測モデルが改善しなかった要因が血圧情報の信頼性であることを示唆している。

要 約

本研究では、DeSCヘルスケア株式会社から受領した大規模リアルワールドデータを用い、腎機能低下予測モデルの開発を試みた。糖尿病治療薬と降圧薬の影響を考慮し、eGFR 45～120 mL/min/1.73 m²の313,974名を対象に機械学習モデルを探索した。結果として、複雑な機械学習モデルとしてXGBoostモデルの予測能が高かったが、最終的に検証コホートで得られたeGFR年間低下度予測モデルの決定係数は4%と極めて限定的であった。相対的重要度が高い特徴量は、ベースライン時のeGFR、年齢、および血圧であった。大迫コホート研究では、家庭血圧がeGFR低下と明瞭に関連したのに対し、健診時血圧との関連は不明瞭だった。このことから、日本の非試験条件下で実施される健診データを用いた検討において、機械学習により予測モデルを改善できなかった要因が、血圧情報の信頼性にある可能性が考えられた。健診データの質向上や家庭血圧の活用など、より信頼性の高いデータ収集方法の検討が今後の機械学習予測モデルの利活用には必要であることが示唆された。

文 献

1. Hashimoto H, Satoh M, Nakayama S, Toyama M, Murakami T, Obara T, Nakaya N, Mori T, Hozawa A, Metoki H、 Comparison of renal prognosis between dipeptidyl peptidase-4 inhibitor users and non-users、 *Diabetes Obes Metab*、 2024
2. Satoh M, Muroya T, Murakami T, Obara T, Asayama K, Ohkubo T, Imai Y, Metoki H、 The impact of clinical inertia on uncontrolled blood pressure in treated hypertension: real-world, longitudinal data from Japan、 *Hypertens Res*、 47 : 598-607、 2024
3. Natsume S, Satoh M, Murakami T, Sasaki M, Metoki H、 The trends of antihypertensive drug prescription based on the Japanese national data throughout the COVID-19 pandemic period、 *Hypertens Res*、 47 : 2086-2090、 2024
4. Umemura S, Arima H, Arima S, Asayama K, Dohi Y, Hirooka Y, Horio T, Hoshide S, Ikeda S, Ishimitsu T, Ito M, Ito S, Iwashima Y, Kai H, Kamide K, Kanno Y, Kashihara N, Kawano Y, Kikuchi T, Kitamura K, Kitazono T, Kohara K, Kudo M, Kumagai H, Matsumura K, Matsuura H, Miura K, Mukoyama M, Nakamura S, Ohkubo T, Ohya Y, Okura T, Rakugi H, Saitoh S, Shibata H, Shimosawa T, Suzuki H, Takahashi S, Tamura K, Tomiyama H, Tsuchihashi T, Ueda S, Uehara Y, Urata H, Hirawa N、 The Japanese Society of Hypertension Guidelines for the Management of Hypertension (JSH 2019)、 *Hypertens Res*、 42 : 1235-1481、 2019
5. Watanabe T, Satoh H, Sugawara Y, Tsukita y, Goto T, Kikuchi A、 The Current State of Blood Pressure Measurement in Medical Examination Institutions –A Study Based on Questionnaire Replies–、 *Health Evaluation and Promotion*、 47 : 463-468、 2020
6. Satoh M, Nakayama S, Toyama M, Hashimoto H, Murakami T, Metoki H、 Usefulness and caveats of real-world data for research on hypertension and its association with cardiovascular or renal disease in Japan、 *Hypertens Res*、 2024