

機械学習を活用した慢性腎臓病患者の 予後予測システムの開発

神田英一郎

川崎医科大学医学部

key words：機械学習，生命予後，末期腎不全，慢性腎臓病

要 旨

慢性腎臓病（CKD）患者は，末期腎不全（ESKD）および死亡のリスクが高い。正確な予測は，高リスク患者のスクリーニングや治療方針の決定に有用である。そこで，CKD 患者における治療の指標として，機械学習モデルによる予測確率を使用できるかどうかを評価し，その実装を試みた。Random Forest, Gradient Boosting Decision Tree, eXtreme Gradient Boosting を使用し，1 年以内のプライマリーアウトカム（ESKD または死亡）の発生を予測するモデルを開発した。CKD 患者のコホートデータを用いて予測精度を評価した（ $n=3,714$ ）。全変数ないし八つの選択した変数を含む二つの RF モデルが，高い予測精度を示した。スプラインを使用した Cox 比例ハザードモデルは，機械学習による予測確率とアウトカム発生の間に統計学的に有意な関係を示した（ $p<0.0001$ ）。さらに臨床で活用するため Web システムを実用化した。本研究は，機械学習に基づく予測確率が ESKD と死亡の新しいリスク指標として有用であることを示した。さらに，Web ベースのリスク予測システムが臨床現場に有用であることが示唆された。

緒 言

慢性腎臓病（chronic kidney disease; CKD）患者は，末期腎疾患（end-stage kidney disease; ESKD），心血管疾患（cardiovascular disease; CVD），および死亡のリスクが高い¹⁻³⁾。さらに chronic kidney disease min-

eral and bone disorder（CKD-MBD），腎性貧血，および高カリウム血症などの合併症も，CKD の進行とともに現れる⁴⁾。CKD の進行を遅らせるには，糖尿病，高血圧，蛋白尿などのさまざまな危険因子を制御する必要がある⁵⁾。

Kidney Disease Improving Global Outcomes（KDIGO）2012 年 CKD ガイドラインは，CKD を診断するための分類システムを提供した⁴⁾。しかし，診療されている CKD の割合は低く^{6,7)}，米国におけるメディケア患者では 14.5% のみであった⁷⁾。潜在的な CKD 患者を特定し，適切な治療が確実に提供されるようにするには，一般の人々の間で CKD とそのリスクの認識を高める必要がある。患者の腎予後と余命の正確な予測は，ESKD のリスクが高い CKD 患者の発見や，治療戦略の決定に有用である^{4,8)}。

ESKD などのアウトカム発生リスクは，統計モデルに基づく臨床指標を使用して予測された確率で評価される⁹⁻¹¹⁾。確率は，リスクの推定，診断，治療など，臨床現場における指標として役立つ¹²⁾。機械学習による予測も確率に基づいている。CKD の進行を評価するためのさまざまな機械学習モデルが開発されてきた¹³⁻¹⁷⁾。しかしながら，これらのモデルの CKD 治療への貢献が期待されているが，いずれも実用化には至っていない。

本研究では，機械学習モデルによって CKD 患者の ESKD または死亡の発生確率を正確に予測できるかどうかを明らかにすることを検証した。さらに，Web ベースのリスク予測システムを開発し，臨床現場での

モデルの実装を試みた。

1 対象・方法

1-1 データ

本研究は、川崎医科大学に通院中の CKD 患者 3,714 人を対象としたコホート研究である (Nos. 3444, 3445, and 5306)。追跡期間は 1 年間であった。患者背景 (年齢, 性別, 合併症) を調査し, 血液検査を施行した。プライマリーエンドポイントは, 透析導入ないし死亡の複合エンドポイントとした。

モデルを開発して選択するために, データを開発用データセット (全患者 80%, $n=2,967$) およびモデル選択用データセットに割り当てた ($n=747$)。開発用データセットの約 70% をモデルのトレーニングに使用し ($n=2,076$)、残りはモデルの検証テストに使用した ($n=891$)。

CKD に関連する因子と薬剤を変数として用いた。外来診療で一般的に使用されており, データ入力が容易であることを考慮して, 変数を選択した。

1-2 統計分析

ベースラインデータは, 必要に応じて, カイ 2 乗検定, t 検定, Mann-Whitney U 検定を使用して比較した。統計的有意性は, 両側 $p<0.05$ として定義した。すべての統計分析は, SAS バージョン 9.4 (SAS Institute, Cary, NC, USA), Python 3.8.2 (Python Software Foundation, DE, USA), および R 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria) を使用した。

1-3 機械学習モデル開発

リスク予測モデルの開発には次の機械学習モデルとして, Random Forest (RF), Gradient Boosting Decision Tree (GB), eXtreme Gradient Boosting (XG) を使用した。機械学習モデルの各タイプには, 変数すべてまたは選択された変数が用いられた。変数の選択は, 各モデル開発時の重要度による順位に基づいて行われた。モデルは, アルゴリズム, データセット, および変数の数によって名付けられた。例えば, x 個の変数を含むベースラインデータに基づく RF モデルは RF_base_xv とし, 時系列データの場合は RF_time_xv とした。

開発されたモデルの予測精度 (C 統計量) は, モデ

ル選択用データに Bootstrap 法を用いて求めた。さらに, サブグループ分析として, 高い糸球体濾過量 (estimated glomerular filtration rate; eGFR) ($60 \text{ mL/min/1.73m}^2$ 以上) および低 eGFR ($60 \text{ mL/min/1.73m}^2$ 未満), 糖尿病 (diabetes mellitus; DM), 非 DM, 若年者 (65 歳未満) および高齢者 (65 歳以上) の各群でモデルの性能を評価した。

十分に開発されたモデルは, 予測される確率が結果のリスクと強く相関するモデルである。モデルに予測された確率とアウトカムリスクの関係を, 予測確率のスプライン項を含む単変量 Cox 比例ハザードモデルでハザード比 (hazard ratio; HR) を評価した。

さらに, ユーザーがデータを入力することで予後予測することができるよう, 得られたモデルをサーバーに搭載し公開した。

2 結果

患者背景を表 1 に示す。65 歳以上の患者数は 1,601 人 (43.1%) であった。CKD ステージの分布は, G1 430 人 (11.6%), G2 804 人 (21.7%); G3a 1,007 人 (27.1%); G3b 660 人 (17.8%); G4 468 人 (12.6%); G5 345 人 (9.3%) であった。モデル開発データセットとモデル選択用データセットに, 患者背景の統計的有意差はなかった。

2-1 機械学習モデルの開発

各モデルでの変数の重要性による順位を図 1 に示す。変数はその重要度によって選択され, 用いられた変数の数は, RF_base では 7, RF_time は 8, GB_base が 13 と 6, GB_time では 7, XG_base は 17 と 7, XG_time が 10 と 6 であった。

図 2 は, 各モデルの 1 年後のプライマリーアウトカム予測の C 統計を示す。いずれのモデルも eGFR や UPCR よりも高い予測精度を示した。特に RF_time_all と RF_time_v8 は他のモデルよりも精度が高く, 0.85 より統計学的有意に高かった ($p<0.05$)。

さまざまな原疾患や背景因子を有する CKD 患者が利用することを考え, サブクラスでモデルの予測能を比較した (表 2)。すべての変数を使用するモデルで, 高い C 統計が得られた。特に, RF_base_all, RF_time_all, RF_time_v8 は高いパフォーマンスを示した。これらのモデルは, 低 eGFR, 糖尿病, 高齢者という

表 1 モデル開発用およびモデル選択用データセット

	モデル開発用データセット	モデル選択用データセット
N	2,967	747
年齢 (歳)	60.1±17.7	60.0±17.4
男 (%)	1,575 (53.1)	396 (53.0)
DM (%)	685 (23.1)	173 (23.2)
高血圧 (%)	1,510 (50.9)	371 (49.7)
CVD (%)	281 (9.5)	82 (11.0)
eGFR (mL/分/1.73m ²)	54.5±31.1	52.8±30.0
アルブミン (g/dL)	3.9±0.7	3.9±0.7
ナトリウム (mmol/L)	140.4±3.0	140.2±3.4
カリウム (mmol/L)	4.4±0.6	4.4±0.6
カルシウム (mg/dL)	9.0±0.7	8.9±0.7
リン (mg/dL)	3.5±0.9	3.5±0.9
LDL (mg/dL)	110.0±35.3	110.6±39.0
尿酸 (mg/dL)	5.9±1.7	5.9±1.7
白血球 (103/ μ L)	6.6±3.3	6.9±3.1
ヘモグロビン (g/dL)	13.0±2.3	12.9±2.2
UPCR (g/gCre)	1.69±3.02 0.49 [0.14, 1.80]	1.71±3.03 0.53 [0.13, 1.74]
RAASI (%)	802 (27.0)	181 (24.3)
ESA (%)	392 (13.2)	116 (15.5)
プライマリーアウトカム (%)	156 (5.3)	34 (4.6)
ESKD (%)	82 (2.8)	21 (2.8)
死亡率 (%)	74 (2.5)	13 (1.7)

連続変数は平均±SD または中央値（四分位範囲）として、カテゴリ変数は n (%) を表す。

Abbreviations:

DM, diabetes mellitus; CVD, cardiovascular disease; eGFR, estimated glomerular filtration rate; LDL, low-density lipoprotein; UPCR, urinary protein-to-creatinine ratio; RAASI, renin-angiotensin-aldosterone system inhibitor; ESA, erythropoietin-stimulating agent; ESKD, end-stage kidney disease.

(著者作成)

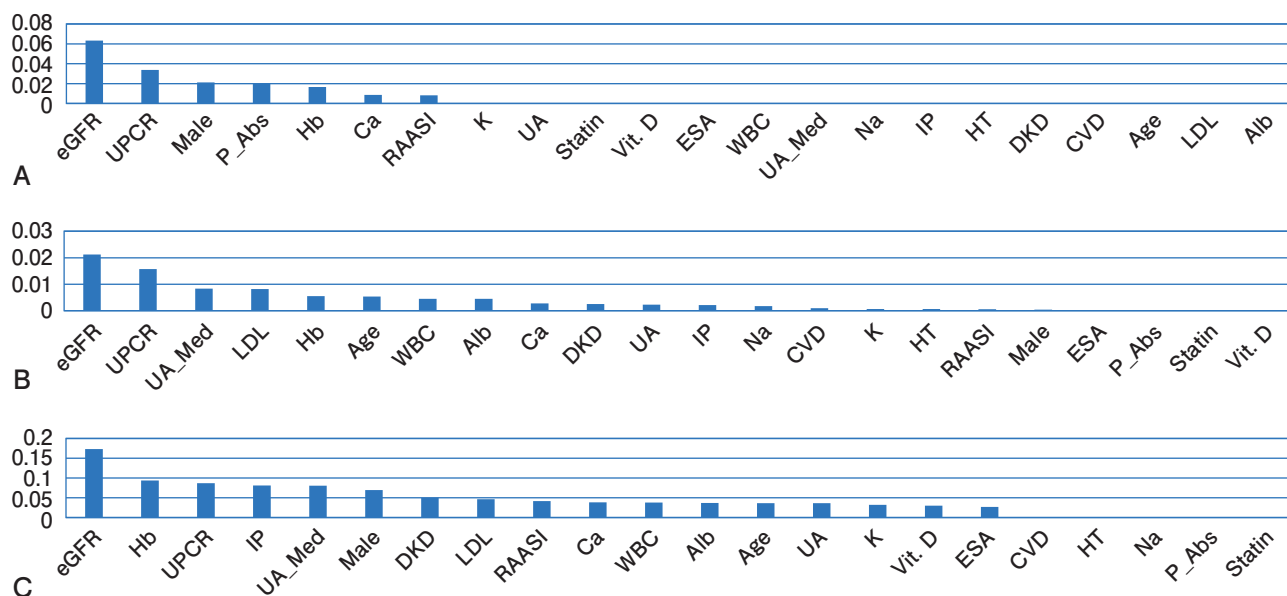


図 1 機械学習モデルにおける変数の重要度

図の数値は各モデルにおける重要度を示す。この重要度により各モデルで使用される変数が選択された。

A. Random Forest

B. Gradient Boosting Decision Tree

C. eXtreme Gradient Boosting

(著者作成)

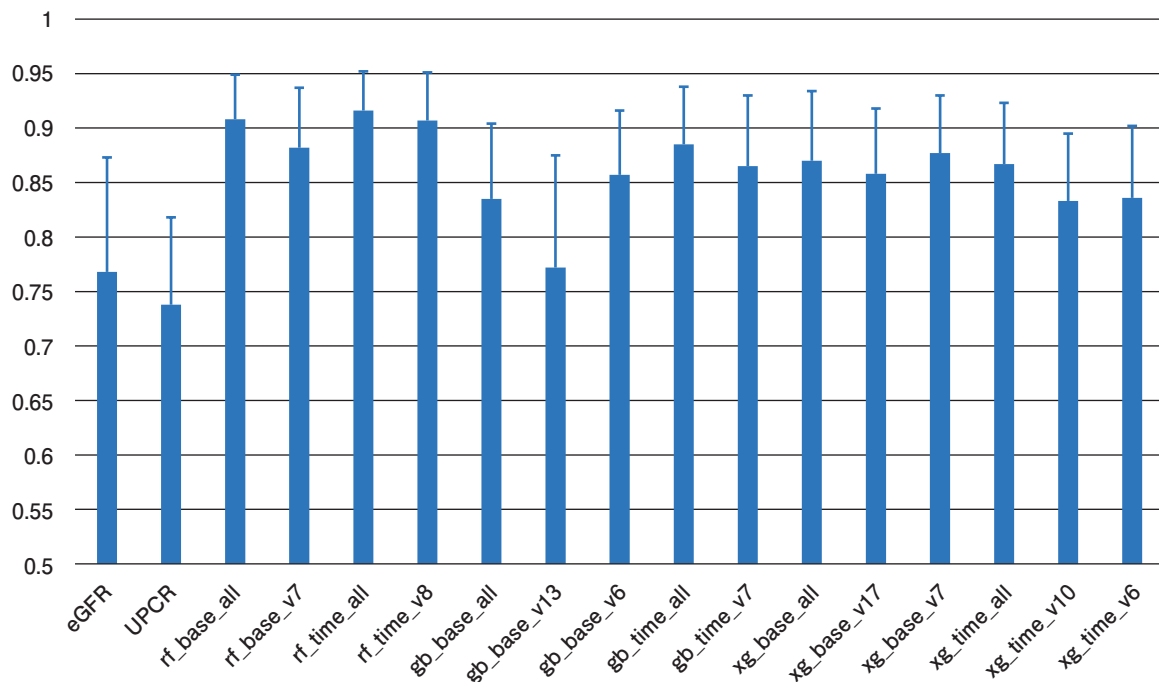


図2 機械学習モデルの予測精度

プライマリーアウトカム発生の予測の精度 (C 統計量と SD)を示す。この数値が高いほど予測精度が高いことを示す。

Abbreviations:

SD, standard deviation; RF, Random Forest; GB, Gradient Boosting Decision Tree; XG, eX-treme Gradient Boosting; base, baseline データセット; time, 時系列データセット; all, 全変数。

(著者作成)

表2 ささまざまな状況下でのモデルの予測能の比較

モデル	プライマリーアウトカム C>0.8	ESKD C>0.8	死亡 C>0.7	eGFR≥60 C>0.75	eGFR<60 C>0.8	Non-DM C>0.9	DM C>0.8	若年者 C>0.9	高齢者 C>0.8	Yes の数
RF_base_all	Yes	Yes	Yes	Yes	Yes	Yes	Yes		Yes	8
RF_base_v7	Yes	Yes	Yes		Yes	Yes	Yes		Yes	7
RF_time_all	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	9
RF_time_v8	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes	8
GB_base_all		Yes	Yes							2
GB_base_v13		Yes								1
GB_base_v6								Yes		1
GB_time_all	Yes	Yes	Yes	Yes		Yes	Yes	Yes		7
GB_time_v7		Yes	Yes				Yes			3
XG_base_all	Yes	Yes			Yes		Yes			4
XG_base_v17		Yes			Yes			Yes		3
XG_base_v7	Yes	Yes	Yes		Yes	Yes				5
XG_time_all	Yes	Yes	Yes	Yes			Yes	Yes		6
XG_time_v10		Yes		Yes						2
XG_time_v6		Yes								1

モデルが要件を満たしている場合、セルに「Yes」が記入される。C>0.7などの表記は、モデルのC統計量がその値よりも統計的に有意に高かったことを示す (p<0.05)。Yesの数が多いモデルは、さまざまなユーザーが使用しても高い予測の性能を示すと考えられる。

Abbreviations:

RF, Random Forest; GB, Gradient Boosting Decision Tree; XG, eXtreme Gradient Boosting; base, baseline データセット; time, 時系列データセット; all, 全変数。(著者作成)

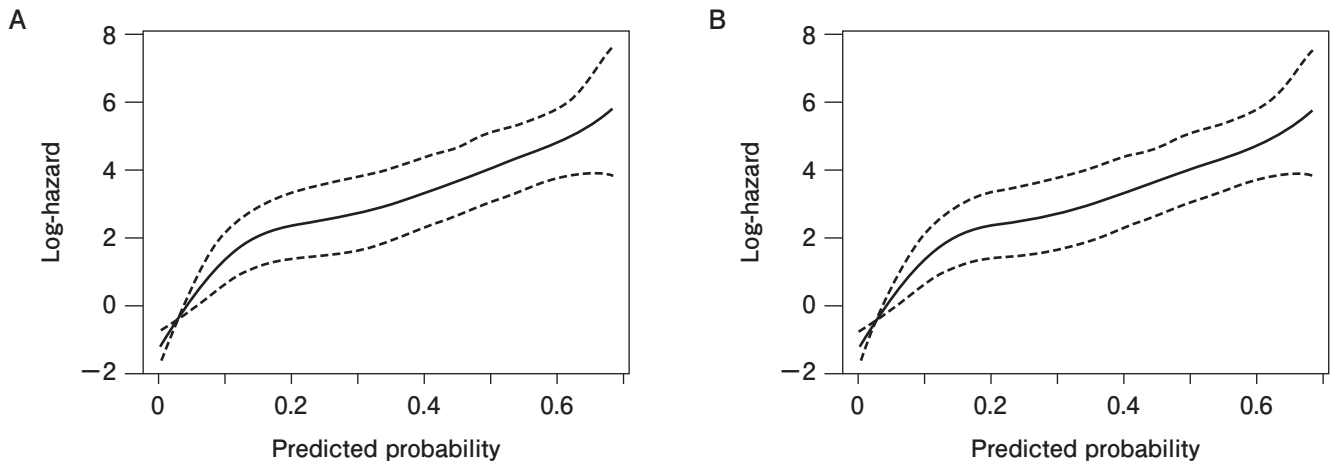


図3 機械学習モデルによる予測確率とアウトカム発生リスクの関係

各機械学習モデルによって予測されたアウトカム発生の確率(X軸)と実際にアウトカムが発生するリスク(Y軸)の関係を示す。リスクはハザード比の対数として表す。

A. RF_time_all

B. RF_time_v8

(著者作成)

CKDの中でも多く認められる患者群でも高性能を示した。これらの結果とサーバーの計算速度を考慮し、Webベースのリスク予測システムには、RF_time_allとRF_time_v8が採用されることになった。

次に、これらのモデルによるプライマリーアウトカム発生の予測確率リスクと実際のイベント発生リスク

の関係を比較した(図3)。RF_time_allとRF_time_v8のHRは、高い予測確率($p < 0.0001$)と関連した。

2-2 予後予測モデルの実装

RF_time_allとRF_time_v8を使用して、Webベースのリスク予測システムを開発した(図4)。このシ

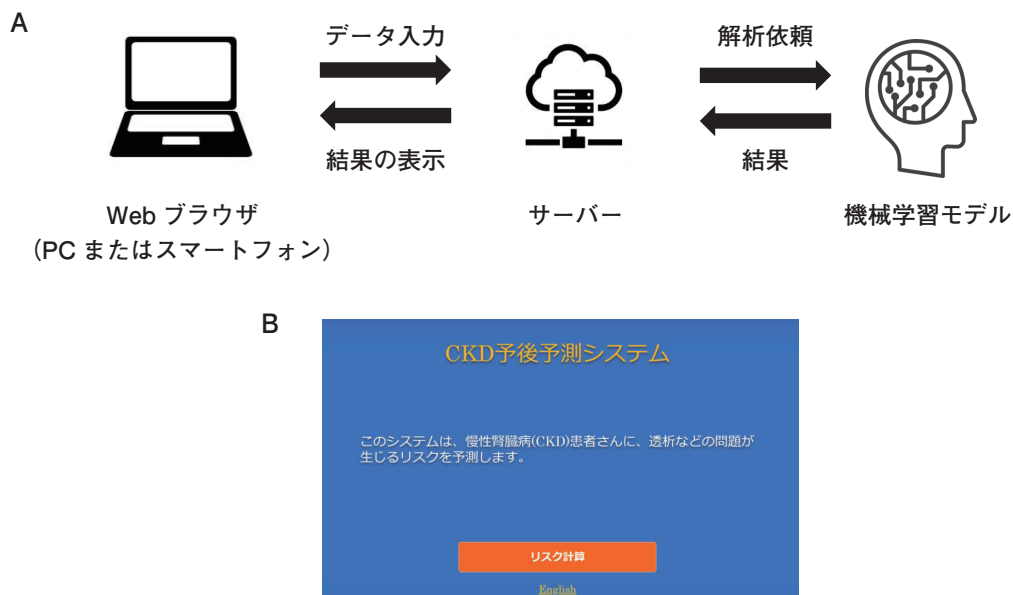


図4 予後予測システムの仕組み

A. 入力されたデータがサーバーを介して機械学習モデルで解析され、その結果が表示されるまでの流れを示す。まず、ユーザーがブラウザからサーバにあるシステムにアクセスする。入力されたデータをサーバに保存されている機械学習で解析する。解析された結果は、ユーザーのブラウザに表示される。

B. 実際の画面。Webサイトに公開した(URL <http://133.242.150.48:8000/> (2023年6月12日現在))。

(著者作成)

システムでは、利用者はPCやスマートフォンのWebブラウザからシステムにアクセスする。次に、このデータはサーバーを介して機械学習モデルに渡され、プライマリーエンドポイントの確率が推定される。最後に、この確率とユーザーへのメッセージがWebブラウザに表示される。容易にアクセスしイベントリスクを簡単に推定できるよう、本システムをWebサイトに公開した (URL <http://133.242.150.48:8000/>)。

3 考察

この研究では、CKD患者の腎臓および生命予後を予測するための機械学習モデルを作成し、このモデルを装備したWebベースのリスク予測システムを開発した。我々の機械学習モデルはさまざまな背景のCKD患者に対しても正確な予後予測をすることができた。本研究の結果から、機械学習モデルによる予測確率は、CKD治療の新しい治療指標となりうることが示された。開発したリスク予測システムは、外来での迅速な予後予測と治療方針の決定を可能にした。本システムはさまざまな患者のデータを用いて徹底的にテストされており、幅広い患者層が利用してもその性能が担保されうると考えられた。

患者の予後の確率は、臨床における医療指標として有用である。CKDの進行を予測する統計モデルを使用して、さまざまなリスク予測モデルが開発されている^{9~11)}。CKD患者の予後予測用の指標にはKidney-failure-risk equation (KFRE)はCox比例ハザードモデルを使用して作成されており、血液透析患者用にはDialysis Outcomes and Practice Patterns Study (DOPPS)がロジスティック回帰モデルを用いてsurvival indexを開発している^{12,18)}。我々のモデルは精度が高く計算しやすいため、実地での活用が期待される。

臨床ではトレードオフを考慮しながら治療計画が選択される。例えば、高リン血症や腎障害の進行を予防するために低たんぱく食を処方すべきか、栄養失調/サルコペニア/フレイルを予防するために低たんぱく食を処方すべきではないのかというように、判断に苦慮する場合がある。機械学習モデルによって推定された確率は、薬剤選択、透析のためのアクセスの決定、および透析の開始など非常に複雑な状況に適用することができ、幅広い用途がある。

多忙な医療スタッフや患者がシステムに迅速かつ簡

単にアクセスできるよう、リスク予測システムをWebサイトにアップロードした。診療の合間に対数関数や指数関数などの指標を多く含む統計式を計算することは、医療関係者にとって非常に困難である。そのため、本システムでは患者データの測定と入力負担を軽減するために、入力データは臨床現場で広く使用されたものを選択し、入力数は少なく制限されている。また、このシステムはOSや電話会社によらないようWebシステムとしており、PCやスマートフォンを使用したインターネット経由での簡単なアクセスを可能にした。このシステムは、AIと情報通信技術の活用によって実現できており、今後さらに使いやすさの向上が期待される。

我々の予後予測システムを臨床現場に活用するには制限がある。まず、日本でデータを取得した。アジアの患者に対する予測精度は高いが、アジア以外の国の患者では結果が偏っている可能性がある。転移学習はグローバルな使用のために、このシステムの精度を向上させるであろう。第二に、ライフスタイル、食物摂取量、家族歴など、数値化して採集しにくいデータが十分に含まれていなかった。第三に、この研究ではRF関連のモデルが主に使用された。ディープラーニングモデルはオーバーフィットする傾向があったため使用されなかった。他のモデルでの検討が今後の課題である。

結 論

本研究は、機械学習で予測された確率は、CKD患者のESKDまたは死亡のリスク増加を正確に反映し、新しいリスク指標として有用であることが示された。さらに、Webベースのシステムが臨床診療に効果的に適用できることが示された。

その他

本研究は令和2年度日本透析医学会公募研究助成によって行われ、成果は原著論文として投稿予定であり、重複掲載を避けるため概要を記した。

利益相反自己申告：利益相反を開示する企業などなし

文 献

- 1) Gansevoort RT, Correa-Rotter R, Hemmelgarn BR, et al. :

- Chronic kidney disease and cardiovascular risk : epidemiology, mechanisms, and prevention. *Lancet* 2013; 382(9889) : 339-52. PubMed PMID: 23727170. Epub 2013/05/31. eng.
- 2) Matsushita K, van der Velde M, Astor BC, et al. : Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts : a collaborative meta-analysis. *Lancet* 2010; 375(9731) : 2073-81. PubMed PMID: 20483451. eng.
- 3) Astor BC, Matsushita K, Gansevoort RT, et al. : Lower estimated glomerular filtration rate and higher albuminuria are associated with mortality and end-stage renal disease. A collaborative meta-analysis of kidney disease population cohorts. *Kidney Int* 2011; 79(12) : 1331-40. PubMed PMID: 21289598. PMCID: PMC3917543. Epub 2011/02/02. eng.
- 4) KDIGO. Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. *Kidney int* 2013; 3 (Supplements 1) : 1-150.
- 5) Inaguma D, Imai E, Takeuchi A, et al. : Risk factors for CKD progression in Japanese patients : findings from the Chronic Kidney Disease Japan Cohort (CKD-JAC) study. *Clin Exp Nephrol* 2017; 21(3) : 446-56. PubMed PMID: 27412450. PMCID: PMC5486452. Epub 2016/07/13. eng.
- 6) Szczech LA, Stewart RC, Su HL, et al. : Primary care detection of chronic kidney disease in adults with type-2 diabetes : the ADD-CKD Study (awareness, detection and drug therapy in type 2 diabetes and chronic kidney disease). *PLoS One* 2014; 9(11) : e110535. PubMed PMID: 25427285. PMCID: PMC4245114. Epub 2014/11/26. eng.
- 7) Saran R, Robinson B, Abbott KC, et al. : US Renal Data System 2019 Annual Data Report : Epidemiology of Kidney Disease in the United States. *Am J Kidney Dis* 2020; 75(1 Suppl 1) : A6-A7. PubMed PMID: 31704083. Epub 2019/11/05. eng.
- 8) 日本腎臓学会編 : エビデンスに基づく CKD 診療ガイドライン = Evidence-based Clinical Practice Guideline for CKD 2018. 東京 : 東京医学社, 2018.
- 9) Echouffo-Tcheugui JB, Kengne AP : Risk models to predict chronic kidney disease and its progression : a systematic review. *PLoS Med* 2012; 9(11) : e1001344. PubMed PMID: 23185136. PMCID: PMC3502517. Epub 2012/11/20. eng.
- 10) Tangri N, Stevens LA, Griffith J, et al. : A predictive model for progression of chronic kidney disease to kidney failure. *JAMA* 2011; 305(15) : 1553-1559. PubMed PMID: 21482743. eng.
- 11) Tangri N, Kitsios GD, Inker LA, et al. : Risk prediction models for patients with chronic kidney disease : a systematic review. *Ann Intern Med* 2013; 158(8) : 596-603. PubMed PMID: 23588748. eng.
- 12) Tangri N, Ferguson T, Komenda P. Pro : Risk scores for chronic kidney disease progression are robust, powerful and ready for implementation. *Nephrol Dial Transplant*. 2017; 32(5) : 748-751. PubMed PMID: 28499025. eng.
- 13) Xiao J, Ding R, Xu X, et al. : Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *J Transl Med*. 2019; 17(1) : 119. PubMed PMID: 30971285. PMCID: PMC6458616. Epub 2019/04/11. eng.
- 14) Liu Y, Zhang Y, Liu D, et al. : Prediction of ESRD in IgA Nephropathy Patients from an Asian Cohort : A Random Forest Model. *Kidney Blood Press Res* 2018; 43(6) : 1852-1864. PubMed PMID: 30537719. Epub 2018/12/07. eng.
- 15) Zhao J, Gu S, McDermaid A : Predicting outcomes of chronic kidney disease from EMR data based on Random Forest Regression. *Math Biosci* 2019; 310 : 24-30. PubMed PMID: 30768948. PMCID: PMC6435377. Epub 2019/02/12. eng.
- 16) Ravizza S, Huschto T, Adamov A, et al. : Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nat Med* 2019; 25(1) : 57-59. PubMed PMID: 30617317. Epub 2019/01/07. eng.
- 17) Kanda E, Kanno Y, Katsukawa F : Identifying progressive CKD from healthy population using Bayesian network and artificial intelligence : A worksite-based cohort study. *Sci Rep* 2019; 9(1) : 5082. PubMed PMID: 30911092. PMCID: PMC6434140. Epub 2019/03/25. eng.
- 18) Kanda E, Bieber BA, Pisoni RL, et al. : Importance of simultaneous evaluation of multiple risk factors for hemodialysis patients' mortality and development of a novel index : dialysis outcomes and practice patterns study. *PLoS One* 2015; 10(6) : e0128652. PubMed PMID: 26030526. PMCID: PMC4451281. Epub 2015/06/01. eng.