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　２ 年　 ５ 月 ２９ 日現在

機関番号：３３９１６
研究種目：研究活動スタート支援
研究期間：2018～2019
課題番号：１８Ｈ０６３８０・１９Ｋ２１４６１

研究課題名（和文）日本人のゲノムデータを活用した正則化回帰モデルによる慢性腎臓病の予測モデル構築

研究課題名（英文）Risk prediction model for chronic kidney disease using regularized regression methods

研究代表者
藤井　亮輔（FUJII, Ryosuke）

藤田医科大学・医療科学部・助教

研究者番号：６０８２３８４６

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：近年は、検査値や生活習慣に加えて、個人のゲノム情報を組み込んだ疾患の予測が注目を集めている。そこで我々は日本人集団約14,000名を対象として、ゲノム情報や基本特性（年齢、性別）などの背景情報をもとに、「解釈が容易かつ精度の高い」を目指し、正則化回帰でモデルと呼ばれる解析手法を用いた腎機能の予測モデルを構築した。これらの手法を使用した場合と従来の解析法である線形回帰モデルとの性能の比較を行った結果、Lasso回帰およびelastic netと呼ばれる手法において、少ない変数でより誤差の少ないモデルを構築できることが推察されたが、性能の向上はわずかであり今後の検討が必要である。

研究成果の学術的意義や社会的意義
本研究では、慢性腎臓病（CKD）をアウトカムとして予測モデルの構築に取り組んだが、他の人種もしくは疾患にも応用可能であり、個人のゲノム情報を使用したリスク予測モデル構築において、大きな可能性を秘めている研究と考えている。また、CKDを経て末期腎不全を発症すると人工透析を要することから、縦断的な検討をさらに実施することで、個人のQOLや社会生産性の向上に貢献しうると考えている。

研究成果の概要（英文）：There has been attracted much interest in the prediction using combined information of individual's genetic variants with laboratory testing value and lifestyle habit. In this study, we developed a predictive model for chronic kidney disease (CKD) using regularized regression models based on genetic information and basic characteristics (age, gender) in a Japanese population (about 14,000 healthy people). Comparing the performance of regularized regression models with that of a conventional analysis method (linear regression model), the Lasso regression and the elastic net may construct a high-performance model with fewer variables and fewer errors. However, the improvement in performance is slight and further studies need to be included more genetic variants.

研究分野：疫学

キーワード：ゲノム疫学　機械学習　正則化回帰モデル　慢性腎臓病　遺伝的多型

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

我が国における慢性腎臓病（CKD）の患者数は、約 1,300 万人と推計されている。CKD が進行し、末期腎不全（ESRD）と診断されると人工透析の治療を受ける必要があり、個人の QOL 低下だけでなく、医療費の増大や生産性の低下など日本の社会的な問題としても注目を集めていた。また、CKD は循環器疾患（CVD）発症のリスク要因であることが日本腎臓学会の診療ガイドラインでも明記されており、CKD の早期発見・発症予防は、CVD 発症の予防にも重要であると考えられている。

疾患の遺伝要因の解明については、次世代シーケンサーなど近年の技術的な進歩により、数十年間で多くの研究者が取り組んできた。特に大規模な集団においては、各遺伝子座と疾患との関係についてゲノム全域を網羅的に探索するゲノムワイド関連解析研究（GWAS）が盛んに行われ、これまで数百個の疾患関連遺伝子が明らかになっていた。とりわけ、欧米人において腎機能をアウトカムとした行なった GWAS によって、CKD と関連する遺伝的な要因が明らかにされていた。このような研究に引き続いて、GWAS の結果をもとに、個人のゲノム情報と生活習慣を活用し、個別化予防につなげる研究が進められていた。ところが、欧米人集団での結果をみると、ゲノム情報を活用したリスク予測の性能は、それらを含まない既存のモデルと同程度であり、さらなる改良が必要であると考えられていた。また、これらの研究成果は欧米人集団のものが大半を占めており、アジア人を含めたその他の人種における検討は遅れをとっていたのが研究当初の背景であった。日本人における当時唯一の GWAS であった菱田らの研究をもとに、申請者が構築した予測モデルの性能も統計学的には有意な性能改善を認めたものの、臨床的にはほとんど意義のないものに留まっていた。そこで、ゲノム情報を含むデータセットに新たな統計手法を適用することで、これまで開発されたモデルよりも精度の高い予測モデル構築に取り組む必要があった。

また、ゲノムデータなどの高次元データ解析においては、既存の統計学的な方法では対応できない課題が複数知られている。一つ目に、N << P 問題である。症例数（N）が変数（この場合は一塩基多型（SNP）の数（P））よりもはるかに少ない状態では、回帰係数の推定精度が大幅に低下する問題がある。二つ目に、結果の解釈性・透明性である。数十万の SNPs を予測モデルに一括投入した場合、その回帰係数を 1 つずつ確認することは困難であり、変数を絞った解釈しやすいモデル構築が求められている。「正則化回帰モデル」は、これらの問題を解決しうる手法として、注目されている機械学習の統計手法の一つである。本手法では、統計モデルを推定すると同時に、膨大な変数から必要な変数を選択する性質を持ち、高次元データを効率的に縮約することができる。医学分野においても、遺伝子の発現量に関するマイクロアレイデータや脳画像処理など膨大な情報を含んだ医学データへの応用も徐々に進んでいた。

２．研究の目的

日本人集団のゲノム情報や検査値などの背景情報を活用し、正則化回帰モデルを適用することにより「解釈が容易かつ精度の高い」CKD 予測モデルを構築する。本研究では、従来の統計手法とは異なる手法を使用してモデルを構築することにより、本邦における CKD 発症の高リスク者を同定する精度向上を目的とする。

３．研究の方法

（１）研究の対象者

日本多施設共同コホート研究（J-MICC Study）のベースライン調査参加者のうち、既にジェノタイピングされた約 14,000 名の日本人集団を対象者とした

（２）遺伝的多型の測定

遺伝的多型の測定は、Illumina OmniExpressExome Array を用いて行われた。SNP 測定後に、質の低い SNP および対象者を除外する品質管理（QC）を行ない、11,184 名の 882,808 SNPs を対象とした。

（３）腎機能の評価

腎機能の評価は、推定糸球体濾過量（eGFR）を用いて行なった。eGFR の推定には、日本腎臓学会が疫学的な研究でも推奨している血清クレアチニン値、年齢、性別から求められる推定式を用いて行なった。また、本研究では、eGFR が 60ml/min/1.73m² 以下に低下している者を CKD と定義した。

（４）統計解析手法

本研究の進め方を図 1 に簡単にまとめた。はじめに、全データを「訓練データ（モデル構築用データセット）」と「検証データ（モデル評価用データセット）」の 2 つに分割する。その後、訓練データにおいて、GWAS を行い関連のある SNP を絞り込んだ後、様々な正則化回帰モデルを適用することで、CKD 予測モデルを構築する。構築したモデルは Receiver operating characteristic（ROC）曲線の area under curve（AUC）や Root mean square error（RMSE）によって評価する。最後に、過適合（over fitting）の問題を避けるために、訓練デ

ータとは独立である検証データにおいて AUC や RMSE を算出する。最終的に、検証データで求めた AUC や RMSE と従来の線形回帰で求めた AUC や RMSE を比較して、正則化回帰モデルで作成した予測モデルの性能および有用性を評価する。
本研究では、それぞれ性質の異なる Ridge, Lasso, elastic net の 3 つの正則化回帰モデルを使用することで、よりデータに適した（精度の高い）モデルが選択できる可能性を広げる。すべてのデータ処理や統計解析は R の glmnet パッケージを用いて実施した。

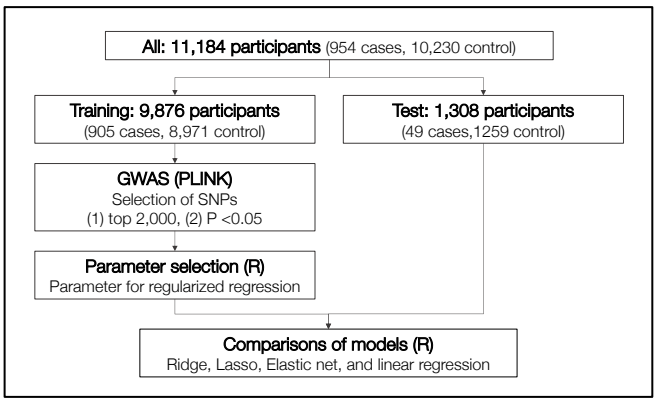

図 1. 研究デザイン

４．研究成果
（１）研究の主な成果
対象集団を表 1 に示す。全体の集団においては、男性が約 46%、平均年齢が 55 歳前後となっており、糖尿病および高血圧の保有率は、それぞれ約 8%、約 37%であった。訓練データと検証データとの間では、検証データの方が基礎疾患の保有率が低かった。この背景の違いによる影響については、後述する。

表 1. 対象者の特性

	全体 (N=11,184)	訓練データ (N=9,876)	検証データ (N=1,308)
男性（人数, %）	5,135 (45.9%)	4,748 (48.1%)	387 (29.6%)
年齢（歳）	54.9 ± 9.3	55.2 ± 9.1	52.7 ± 10.4
eGFR (mL/min/1.73m^2)	78.7 ± 15.1	77.9 ± 14.8	84.3 ± 15.7
高血圧（人数, %）	4,171 (37.3%)	3,835 (38.8%)	336 (25.6%)
糖尿病（人数, %）	921 (8.2%)	872 (8.8%)	49 (3.7%)
CKD (人数, %)	954 (8.5%)	905 (9.2%)	49 (3.7%)

訓練データにおける予測モデルのパラメータを決める段階で、SNP を絞り込む基準を 2 つ設けて研究を進めた。1 つ目は、GWAS を行ない Top 2,000 SNP を使用するものであり、2 つ目は、p 値が 0.05 を下回る SNP を使用するものである。それぞれのシナリオにおける結果を表 2 に示す。

表 2. 各正則化回帰モデルで使用した変数の数と RMSE、R^2

Methods	変数の数	RMSE	R^2 (95%CI)
#1. Top 2,000 SNPs			
線形回帰	2,009	0.209	0.228 (0.176-0.279)
Ridge 回帰	2,009	0.192	0.275 (0.226-0.326)
Lasso 回帰	863	0.186	0.348 (0.299-0.394)
Elastic net	963	0.187	0.341 (0.292-0.388)
#2. SNPs of p-value < 0.05			
線形回帰	31,540	0.289	0.070 (0.019-0.126)
Ridge 回帰	31,540	0.203	0.131 (0.078-0.184)
Lasso 回帰	2,985	0.190	0.323 (0.274-0.371)
Elastic net	1,753	0.185	0.362 (0.314-0.408)

このように、1 つ目のシナリオにおいて、Lasso 回帰および elastic net で変数を絞り込んだ場合、約 800～900 SNP を使用して、RMSE が 0.185 前後、R^2 が 0.340 前後であった。一方で、全部の変数を使用している線形回帰や Ridge 回帰では、RMSE が 0.190～0.200 程度、R^2 が 0.230～0.275 であった。また、2 つ目のシナリオについても同様に、全部の変数を使用している線形回帰モデルや Ridge 回帰よりも変数選択の特徴を持つ Lasso 回帰および elastic net において RMSE が小さく、R^2 が大きい結果を得た。
対象者の特性にやや偏りが見られることやより汎化性能を向上させるために、さらなる追加解析を実施した。まずは、完全にランダムな状態で訓練データと検証データを振り分けるランダムサンプリングを行なったところ、最も精度が良いモデルは、897 個の SNP を使用した elastic net（RMSE=0.200）であり、その次に 826 個の SNP を使用した Lasso（RMSE=0.201）であった。通常のロジスティック回帰モデルでは、2009 個の SNP を使用して、RMSE=0.216 であった。次に、ランダムサンプリングに加えて、最初から最後までのプロセスを 5 回繰り返す解析では、平均して 804 個、898 個の SNP を使用した Lasso と elastic net どちらも平均 RMSE が 0.196 と最も小さい値であった。

（2）結果の考察と今後の展望
本研究において、Lasso 回帰および elastic net の正則化回帰モデルでは、変数を絞った上で性能（RMSE および R^2）を改善することが証明された。一方で、わずかな性能向上の臨床的な意義は定かでないことや、近年は全ての SNP を使用した方が予測能が高いという研究成果が示されていることから、変数選択の性質を持っている正則化回帰モデルのゲノム情報に適用することについては再考する必要があると考えられる。今後の展望としては、1）より効果の明らかになっている SNP を使用する、2）遺伝子と生活習慣の遺伝的相互作用、3）頻度の低い SNP（MAF<5%）などの今回含んでいない条件等を加えることでさらに予測能の向上を図っていく。

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　2件）

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 藤井亮輔
２．発表標題 日本人集団におけるメンデルランダム化解析による高感度CRPとeGFRとの関連：J-MICC STUDY
３．学会等名 令和元年度 コホート・生体試料支援プラットフォーム 若手支援研究成果発表会
４．発表年 2020年

１．発表者名 Ryosuke Fujii, Asahi Hishida, Takeshi Nishiyama, Masahiro Nakatochi, Takaaki Kondo, Kenji Wakai, J-MICC Study group
２．発表標題 High-sensitivity C-reactive protein and estimated glomerular filtration rate: a two-sample mendelian randomization
３．学会等名 The 69th Annual Meeting of the American Society of Human Genetics（国際学会）
４．発表年 2019年

１．発表者名 藤井亮輔，近藤高明，菱田朝陽，大西丈二，内藤真理子，浜島信之，若井建志
２．発表標題 飲料品の摂取量と腎機能低下との関連：J-MICC Study大幸地区
３．学会等名 第78回日本公衆衛生学会総会
４．発表年 2019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