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　５ 年　 ６ 月 ２１ 日現在

機関番号：１７３０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１１７２３

研究課題名（和文）健康医療データと全ゲノム情報間の相互作用を考慮したリスク予測モデリング

研究課題名（英文）Risk prediction modeling by accounting for interaction between health-related data and whole-genome information

研究代表者
植木　優夫（Ueki, Masao）

長崎大学・情報データ科学部・教授

研究者番号：１０５１５８６０

交付決定額（研究期間全体）：（直接経費）　　1,800,000円

研究成果の概要（和文）：全ゲノム情報に多様な健康医療データを組み合わせたリスク予測モデリングのための統計手法・アルゴリズムの開発、特に、健康医療データと全ゲノム情報間の相互作用を考慮したモデリング手法の開発を行った。スパースモデリング手法STMGP（smooth-threshold multivariate genetic prediction）を用いて、全ゲノムデータと性別や年齢等の背景因子を組み合わせたリスク予測モデリング手法の評価を実際のデータで行った後に、全ゲノムデータと非ゲノムデータの相互作用を遺伝子×環境相互作用効果としてモデルに組み入れられる非相加的モデルに基づくリスク予測モデリング手法を開発した。

研究成果の学術的意義や社会的意義
近年、ゲノムデータを含め、高次元な健康医療データが取得されているが、十分な疾患リスク予測精度が得られないケースが多くある。本研究において、現行の単純な加法モデルを発展させることで、全ゲノム情報と多様な健康医療データの相互作用を考慮できる非線形リスク予測モデルを開発した。これまでゲノムデータに対する予測モデルにおいて非ゲノムデータとの相互作用を考慮できる予測モデルは限られていたが、本手法を用いることで、ゲノムデータと健康医療データの相互作用が存在する場合の予測精度向上に貢献するものと考える。

研究成果の概要（英文）：Combining whole-genome data with various health-related data, we developed statistical models and algorithms for risk prediction. In particular, we developed prediction model that incorporates interactions between genome-data and health-related data. Based on the STMGP (smooth-threshold multivariate genetic prediction), a sparse modeling method, we evaluated the prediction model combing whole-genome data and other factors such as sex and age on real dataset. Subsequently, we developed a prediction model that incorporates whole-genome data, non-genomic data (sex, age, etc), and their interactions, which is a non-additive gene-environment interaction based prediction model.

研究分野：統計学

キーワード：予測モデル　遺伝子x環境相互作用　高次元回帰モデル　健康医療データ　ゲノムデータ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1) 個人の背景に応じた将来の疾患リスクを高精度に見積もれるようになれば、精密医療の実現に向けた大きな一歩となる。研究開始時点において、個人の全ゲノム情報およびその他の因子（性別、年齢、環境要因など）から、疾患発症等のリスクを統計的に予測する方法が着目されていた。しかし全ゲノム情報を用いた予測モデルの予測精度は、多くのありふれた疾患において実用水準に達していないことが問題視されていた。データ収集環境の発展に伴い大量に収集されはじめていた生活習慣や健康診断情報などの様々な健康医療データを、全ゲノム情報と組み合わせて用いることで、現行モデルの予測精度を向上させられる可能性を予想した。

(2) 手法面では、特にサンプルサイズが十分でない状況において、多数の要因を含む高次元データに対する予測モデリング手法は未だ発展途上にあった。このことが十分な性能が得られない理由のひとつであると考え、方法論的な側面からモデリングを発展させることで、リスク予測の精度を向上できる可能性を期待した。ゲノムデータに対する予測モデリング手法においては、ゲノムデータの大規模さと超高次元性により、ゲノム情報の加法モデルが広く用いられていた。他方で、健康医療データなどの非ゲノム情報は、ゲノム情報と独立に用いられるか、もしくは、ごく少数の非ゲノム情報がゲノムデータに加法モデルの枠組み内で同列に扱われることが通常であった。したがって、健康医療データと全ゲノム情報間の相互作用が考慮されておらずモデリングが限定的であったため、これらの間の相互作用を考慮した非加法モデルを開発することによりリスク予測の高精度化を目指した。

２．研究の目的

全ゲノム情報に多様な健康医療データを柔軟に組み合わせて、ゲノム情報と健康医療データを加法モデルで扱う現行のリスク予測モデルよりも高い予測精度を示す統計手法・アルゴリズムの開発を目的とした。

３．研究の方法

(1) 様々な健康医療データと全ゲノム情報を取り扱う統計手法として、ごく単純な加法モデルに基づく統計手法が広く使われていたが、それらを組み合わせて解析するための統計手法や方法論が整備されていなかったため、方法論・理論的な研究から取り組んだ。健康医療データと全ゲノム情報と非加法的に組み合わせる手法として、両者の相互作用を考慮した解析手法について研究を行った。また健康医療データおよび全ゲノム情報の双方ともに高次元データであるため、並行して高次元データに対する予測手法についても調査し、理論面を含めて研究を進めた。特に、候補の変数の数がサンプルサイズよりも遥かに大きい場合である超高次元データで動作する高次元回帰手法の研究を行った。さらに、変数間の相関関係を取り扱う方法、相互作用の組み合わせ数増大に伴う計算実行可能性への対処、複雑な健康医療データのモデリング手法、について研究を進めた。

(2) 統計手法を数理的に研究し、理論的性質を解析した後で、アルゴリズムを開発し、プログラム実装を行った。また、実際のデータを模倣した大規模数値実験を行って、有限サンプルでの性能評価と検証を行った。さらに、実際のデータ（公共データを利用）を通じて開発した統計手法とアルゴリズムの性能を評価した。

４．研究成果

(1) これまで研究代表者が開発したスパースモデリング手法 STMGP（smooth-threshold multivariate genetic prediction、引用文献①）を用いて、全ゲノムデータと性別や年齢等の背景因子を組み合わせたリスク予測モデリングをうつ症状のリスク予測に適用し、予測性能を評価した。大規模ゲノムデータ解析のための計算機実装を行い、実際のゲノムデータを模倣したシミュレーション研究、実際のうつ症状データに対する予測性能の検証を通じて、STMGP 法が既存手法（ポリジェニックリスクスコア、ゲノミック BLUP、BayesR 等）を上回る予測性能を示すことを見出した（引用文献③）。

(2) 数百種類の代謝物データを用いたうつ症状に対する予測モデリングの研究を行った HSIC Lasso、Lasso、サポートベクターマシン、PLS、ランダムフォレスト、ニューラルネットワークの予測性能を比較し、HSIC Lasso が高い性能を示すことを見出した（引用文献②）。

(3) STMGP 法（引用文献①）のアルゴリズムを拡張し、全ゲノムデータと非ゲノムデータ（性別や年齢等の背景因子）、さらにはそれらの相互作用をモデルに組み入れることのできる遺伝子×環境相互作用を考慮した非相加的モデルに基づくリスク予測モデリング手法を開発した。STMGP 法では各ゲノムデータに対して 1 変量回帰を繰り返して得られる P 値を利用していたが、今回開発したリスク予測手法では、遺伝子×環境相互作用解析を繰り返して得られる P 値を用いた。遺伝子×環境相互作用モデルでは推定対象のパラメータが複数あるが、これを 1 変量回帰で近似できる新たな近似統計量を開発し、1 変量回帰の枠組みに落とし込んで 1 変量回帰の P 値に持ち込むことで STMGP 法を直接適用可能とした。遺伝子×環境相互作用を仮定したシミュレーション実験を通じて、開発した手法が既存の相加的モデルを上回る性能を示すことを確認した。また、実際のゲノムデータに適用したところ、良好なリスク予測性能を示した（引用文献⑤）。

(4) 高次元回帰手法を仮説検定に利用するための統計的手法を新たに開発し、Lasso や Elastic Net などの罰則付き回帰モデルを用いて、ゲノムデータからの遺伝的要因探索に用いることのできる統計手法を開発した。ゲノムデータの解析においては、1 変量回帰における検定を繰り返して得られる P 値を用いたスクリーニングが行われる。STMGP 法でもこの 1 変量回帰の P 値を用いるが、1 変量回帰が独立に適用されるため遺伝子間の相関が考慮されない。相関を考慮した検定が望ましいが、高次元データに対しては容易でない。遺伝子などの事前に定義された変数グループを利用できる場合においては，グループ単位での解析方法が提案されているものの、グループの定義自体が不確かさを含んでいる場合は検出力が限定的となる。そこで、モデル選択を取り入れたデータ適応的検定手法を提案し、例えば罰則付き回帰のように複雑度の低いモデルから複雑度の高いモデルを繋ぐ回帰モデル列からモデルを選択した後で条件付き期待値に対する検定を行う手法を開発した。本手法は、柳井の一般化決定係数を用いた検定を利用することで、確率値を算出する際にシミュレーションによる高負荷な計算が不要であるため、ゲノムデータなどの大規模な検定が必要とされる状況であっても容易に使用できる。様々な条件で行った数値実験により、提案法がタイプ１エラーを制御しつつ高い検出力が得られることを確認した。さらに、実際のゲノムデータに適用し、数値実験と同様の良好な結果が示された（引用文献④）。

(5)新型コロナウイルス感染症の感染者数の時空間データに対して時空間データモデリングを行うため、非線形ランダム効果とベータ負の二項分布を用いた新たな手法を開発した。本手法は、時空間データにおける相関構造をモデル化し、感染者数の時空間的な予測モデリングを行うことができる。既存の負の二項分布を用いる同様のモデルを実際の国内の新型コロナ感染症の感染者数の時空間データに適用したところ、データの急激な変化を十分に捉えられなかったが、開発した手法はこの変化を捉えることができた。これは、極端なデータをモデル化できるベータ負の二項分布の特性に由来する。さらに、開発した手法は予測性能の面でも良好な結果を示した（引用文献⑥）。

＜引用文献＞

① Ueki M, Tamiya G. Smooth-threshold multivariate genetic prediction with unbiased model selection. Genetic Epidemiology 40: 233-243, 2016.
② Takahashi, Y., Ueki, M., Yamada, M. et al. Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection. Translational Psychiatry 10:157, 2020.
③ Takahashi, Y., Ueki, M., Tamiya, G. et al. Machine learning for effectively avoiding overfitting is a crucial strategy for the genetic prediction of polygenic psychiatric phenotypes. Translational Psychiatry 10:294, 2020.
④ Ueki M. Testing conditional mean through regression model sequence using Yanai's generalized coefficient of determination. Computational Statistics & Data Analysis 158:107168, 2021.
⑤ Ueki M, Tamiya G. Smooth-threshold multivariate genetic prediction incorporating gene-environment interactions. G3 Genes | Genomes | Genetics 11: jkab278, 2021.
⑥ Ueki M. Beta-negative binomial nonlinear spatio-temporal random effects modeling of COVID-19 case counts in Japan. Journal of Applied Statistics 2022.

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　4件）

１．著者名 Ueki M	４．巻 158
２．論文標題 Testing conditional mean through regression model sequence using Yanai's generalized coefficient of determination	５．発行年 2021年
３．雑誌名 Computational Statistics & Data Analysis	６．最初と最後の頁 107168～107168
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.csda.2021.107168	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 Ueki M, Tamiya G	４．巻 11
２．論文標題 Smooth-threshold multivariate genetic prediction incorporating gene-environment interactions	５．発行年 2021年
３．雑誌名 G3 Genes|Genomes|Genetics	６．最初と最後の頁 jkab278
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/g3journal/jkab278	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 Takahashi, Y., Ueki, M., Yamada, M. et al.	４．巻 10
２．論文標題 Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection	５．発行年 2020年
３．雑誌名 Translational Psychiatry	６．最初と最後の頁 157
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41398-020-0831-9	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 Narita, A., Nagai, M., Mizuno, S. et al.	４．巻 10
２．論文標題 Clustering by phenotype and genome-wide association study in autism	５．発行年 2020年
３．雑誌名 Translational Psychiatry	６．最初と最後の頁 290
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41398-020-00951-x	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 Takahashi, Y., Ueki, M., Tamiya, G. et al.	4．巻 10
2．論文標題 Machine learning for effectively avoiding overfitting is a crucial strategy for the genetic prediction of polygenic psychiatric phenotypes	5．発行年 2020年
3．雑誌名 Translational Psychiatry	6．最初と最後の頁 294
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41398-020-00957-5	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 Narita Akira、Ueki Masao、Tamiya Gen	4．巻 66
2．論文標題 Artificial intelligence powered statistical genetics in biobanks	5．発行年 2020年
3．雑誌名 Journal of Human Genetics	6．最初と最後の頁 61～65
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s10038-020-0822-y	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3件（うち招待講演　1件／うち国際学会　2件）

1．発表者名 Ueki M
2．発表標題 Data-adaptive groupwise test for genomic studies via the Yanai's generalized coefficient of determination
3．学会等名 Bernoulli-IMS 10th World Congress in Probability and Statistics（国際学会）
4．発表年 2021年

1．発表者名 植木優夫
2．発表標題 データ科学による遺伝統計解析
3．学会等名 脳病態数理・データ科学セミナー（招待講演）
4．発表年 2021年

１．発表者名 Masao Ueki, Gen Tamiya
２．発表標題 Sparse genetic prediction modeling incorporating gene-environment interactions
３．学会等名 30th International Biometric Conference (IBC2020)（国際学会）
４．発表年 2020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

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