東北大学
TOHOKU UNIVERSITY

Press Release

2020 年 8 月 12 日

報道機関　各位

東北大学東北メディカル・メガバンク機構
岩手医科大学いわて東北メディカル・メガバンク機構
国立研究開発法人　日本医療研究開発機構

機械学習を用いたうつ病症状のリスク予測の研究
～精神疾患の個別化医療を目指して～

【発表のポイント】

- 精神疾患のリスクを説明するような遺伝子情報を適切に組み合わせる数理モデルがないことが問題となっていました。
- 複数の数理モデルを比較し、過学習[*1]を抑えるような機械学習手法を用いた手法が、うつ病症状をはじめとする精神疾患のリスク予測に有用なことを示しました。
- 今回有効とされた手法を用いて精密な疾患のリスク予測が可能となれば、うつ病に関する個別化医療や予防、遺伝素因に関する病態生理の解明に寄与することが期待されます。

【概要】

うつ病の発症には多数の DNA 多型が関係すると想定されていますが、関係する遺伝子個々の影響はとても小さいことがわかっています。そうした小さい効果がどのように組み合わされてうつ病のなりやすさ（脆弱性）が形成されるのかはまだよくわかっていません。東北大学東北メディカル・メガバンク機構の高橋雄太医員、植木優夫助教（現・長崎大学教授）、田宮元教授、富田博秋教授らは、うつ病症状に関連する DNA 多型情報[*2]について機械学習手法を用いたこれまでの研究で個別化医療につながる知見を得ました。

今回の研究では、先に開発した機械学習手法である STMGP 法[*3]を用いてうつ病に関する症状や種々のシミュレーションデータを使った解析を行うことで、多数の DNA 多型が複雑に病態に関係していることが想定されるうつ病をはじめとする精神疾患のリスク予測に、STMGP 法が有用であることが示唆されました。

本研究は、日本医療研究開発機構（AMED）の脳科学研究戦略推進プログラムにおける課題「栄養・生活習慣・炎症に着目したうつ病の発症要因解明と個別化医療技術開発」によって行われました。
この成果は米国時間 2020 年 8 月 17 日に米国科学雑誌「Translational Psychiatry」のオンライン版で公開されます。

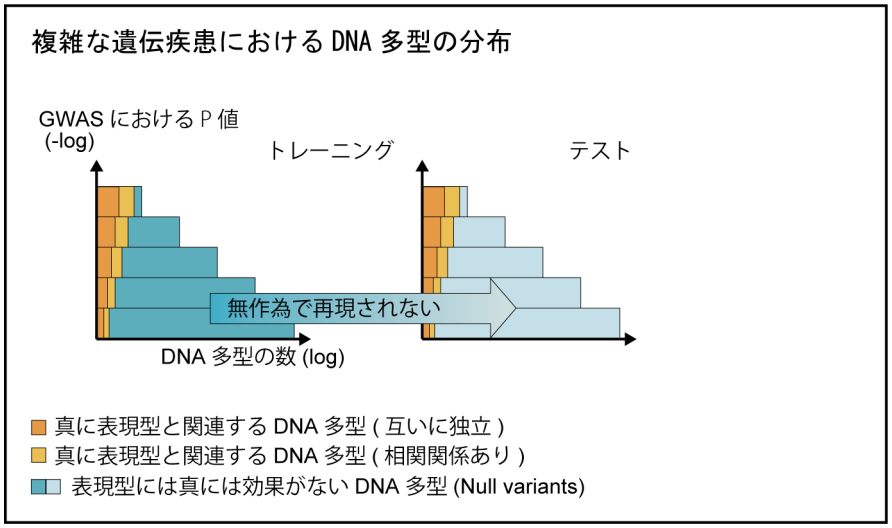
【詳細】

精神疾患の発症には遺伝子をコードしている DNA 上に存在して個体差を規定する DNA 多型が関わることが以前より知られていましたが、精神疾患の発症を DNA 多型情報から予測モデルを用いてリスク予測する研究には大きな課題がありました。それは多くの精神疾患の発症には多数の DNA 多型が関わっている一方、一つ一つの DNA 多型の効果サイズが小さいため、真に疾患との関連を示す DNA 多型を検出することが難しく、結果的に予測モデルの過学習が生じることとなり、予測精度の向上は限られてしまうからです。STMGP 法は DNA 多型の選択をして、さらに GWAS[*4] の統計量に基づいて DNA 多型に重みづけを行うことで予測モデルの過学習を抑え、高精度な予測ができるように開発された機械学習手法です（図）。今回の研究において、この STMGP 法が実際に精神疾患発症のリスク予測をするのに有効であるかを検討しました。

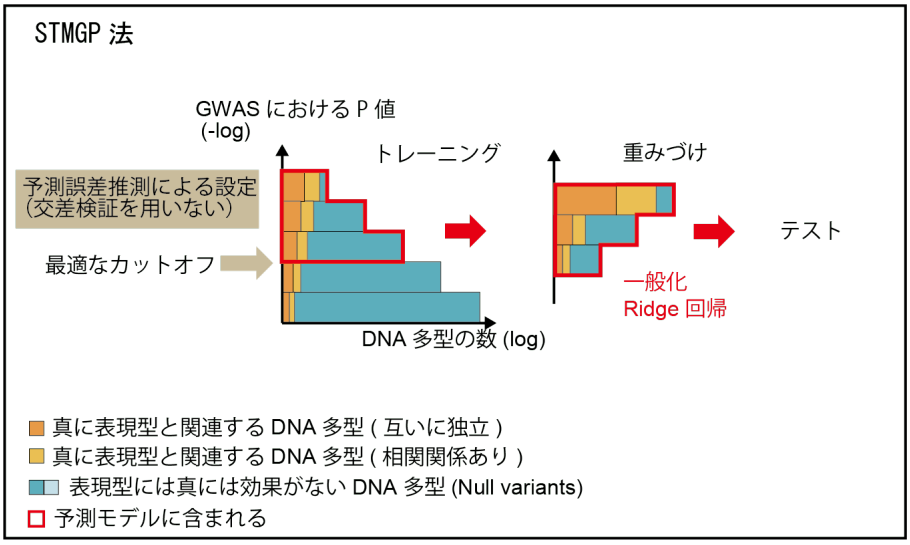
具体的には、東北メディカル・メガバンク計画によって収集された宮城県在住の 3,685 人分の DNA 多型情報を用いて予測モデルの機械学習を行い、岩手県在住の 3,048 人のデータを用いて予測モデルの精度を評価しました。まず、うつ病に関する症状のデータと、さまざまなシミュレーション解析[*5] により作成されたデータを用いて予測精度の検討を行いました。次に、STMGP 法と現時点で頻用されている最先端の予測モデル（Polygenic Risk Score 法、genomic best linear unbiased prediction, summary-data-based best linear unbiased prediction, Bayes R 法、Ridge 回帰法）について、それぞれの予測精度と過学習の程度を比較しました。

これらの比較検討の結果、うつ病に関する症状のリスク予測においては、他モデルの予測精度と比較して有意差はないものの、STMGP 法は最も高い予測精度を示しました。シミュレーション解析の検討においては、STMGP 法が複雑な遺伝素因をもつ精神疾患の DNA 多型情報からのリスク予測に有用である可能性を示しました。

今後研究が進み、STMGP 法といった過学習を抑えた機械学習の手法を用いることで精密なうつ病のリスク予測が可能となれば、うつ病やその他の精神疾患への罹患のし易さやし難さに関わる病態の一部を DNA 多型に基づいて説明でき、より個々人に適した予防法や医療を提供することが期待されます。



【図 1】



【図 2】

図 1 に示すように、精神疾患をはじめとする多数の DNA 多型情報が複雑に関与する疾患においては、ゲノムワイド関連解析（GWAS）における P 値[*6]のみで DNA 多型を選別しても、検出力が不足しているために、予測モデルのトレーニングデータとテストデータで結果が再現されないことが多いとされています。STMGP 法では図 2 に示すように、予測誤差[*7]推測により P 値のカットオフを定め、GWAS の統計量も用いた重みづけを一般化 Ridge 回帰[*8]という機械学習手法で行います。これにより、過学習を抑えながらのリスク予測が可能となります。

【論文名】

Machine learning for effectively avoiding overfitting is a crucial strategy for the genetic prediction of polygenic psychiatric phenotypes

「機械学習を用いて過学習を抑えることが、複雑な遺伝素因をもつ精神疾患のリスク予測には重要である」

Yuta Takahashi, Masao Ueki, Gen Tamiya, Soichi Ogishima, Kengo Kinoshita, Atsushi Hozawa, Naoko Minegishi, Fuji Nagami, Kentaro Fukumoto, Kotaro Otsuka, Kozo Tanno, Kiyomi Sakata, Atsushi Shimizu, Makoto Sasaki, Kenji Sobue, Shigeo Kure, Masayuki Yamamoto, Hiroaki Tomita

掲載誌：Translational Psychiatry

DOI：10.1038/s41398-020-00957-5

【参考】

＜機械学習を用いた精神疾患のリスク予測研究について＞

今回の研究チームは生物学的な指標に機械学習を用いることで精神疾患との関連を示し、個別化医療につなげることを目指しています。今回の研究の他に、理化学研究所革新知能統合研究センター（AIP）の山田誠チームリーダー（現・京都大学）らとの共同研究で、東北メディカル・メガバンク計画によって収集された血漿中の代謝物情報に機械学習を用いることで精神疾患との関連を探った下記の研究が今年 5 月に公開されています。

Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection

「非線形特徴量選択機械学習を用いることでうつ病関連症状の代謝物からの予測精度は向上する」

Yuta Takahashi, Masao Ueki, Makoto Yamada, Gen Tamiya, Ikuko N. Motoike, Daisuke Saigusa, Miyuki Sakurai, Fuji Nagami, Soichi Ogishima, Seizo Koshiba, Kengo Kinoshita, Masayuki Yamamoto, Hiroaki Tomita

掲載誌：Translational Psychiatry

DOI：10.1038/s41398-020-0831-9

＜東北メディカル・メガバンク計画について＞

東北メディカル・メガバンク計画は、東日本大震災からの復興と、個別化予防・医療の実現を目指しています。東北大学東北メディカル・メガバンク機構（ToMMo）と岩手医科大学いわて東北メディカル・メガバンク機構（IMM）を実施機関として、東日本大震災被災地の医療の創造的復興及び被災者の健康増進に役立てるために、平成 25 年より合計 15 万人規模の地域住民コホート調査および三世代コホート調査等を実施して、試料・情報を収集したバイオバ

ンク[*9]を整備しています。本計画については、平成 27 年度より、AMED が研究支援担当機関の役割を果たしています。

【用語説明】

＊1.　過学習：一般に予測モデルを作成する場合、予測モデルを学習させるためのデータ(トレーニングデータ)とそのモデルの性能を評価するためのデータ(テストデータ)の 2 つを用意する。過学習とは、学習の段階ではあたかも性能が良いかのように高い予測精度を示すが、実際のテストの段階では予測精度が低くなってしまうことを指す。

＊2.　DNA 多型情報：人体を形作るタンパク質などの構成分子の設計図となる遺伝子はアデニン、シトシン、グアニン、チミンの 4 つの塩基が 30 億つながるゲノム DNA 上にコードされている。この塩基の配列の千塩基に 1 つは個人ごとに異なる塩基からなる箇所があり、DNA 多型と呼ばれる。この DNA 多型を検出して集約した情報を指す。

＊3.　STMGP 法（smooth-thresholded multivariate genetic prediction 法）：ToMMo 田宮元教授らのグループが開発した複雑な遺伝疾患のリスク予測を高精度に可能とする機械学習手法。（植木ら、2016 年、DOI: 10.1002/gepi.21958.）

＊4.　GWAS（Genome-Wide Association Study 、ゲノムワイド関連解析）：ヒトゲノムの全体をほぼカバーする数百万から数千万の DNA 多型情報について、形質と合わせて統計学的な処理を行うことで DNA 多型と形質の関連性を調べる解析手法。

＊5.　シミュレーション解析：ここではゲノムワイド DNA 多型情報（GWAS）をもとに仮想上のスコアを作成し、この予測精度を調べることを指す。仮想上のスコアは、それぞれの DNA 多型に確率・統計学的な解析を行い作成したもの。ここでシミュレーション解析を行う目的は、いろいろな遺伝的な構造をもつ表現型を仮想的に多数作り出して、それぞれに対する STMGP の予測精度を調べ、うつ病症状以外にも STMGP が有効なものがあるかを探すことである。

＊6.　P 値：統計的仮説検定において、「その DNA 多型が疾患と関連しない」という帰無仮説が棄却される確率のこと。図では P 値の負の対数（−log）が Y 軸に示してあるため、図の上に示した領域の多型の方が、より小さい P 値を示す多型、すなわち、統計的に関連が示唆される多型であることを意味している。

＊7.　予測誤差：予測値と実績値の差

＊8.　Ridge 回帰：非常に多くの変数の値をもとに一つの変数を予測する場合に、過学習を抑えることで予測精度を上げる機械学習手法の一つ。STMGP

法では GWAS の統計量による重みづけも行う「一般化」Ridge 回帰を使用している。

＊9.　バイオバンク：生体試料を収集・保管し、研究利用のために提供を行う。東北メディカル・メガバンク計画のバイオバンクは、コホート調査の参加者から血液・尿などの生体試料を集める。

【お問い合わせ先】
（研究に関すること）
東北大学東北メディカル・メガバンク機構
脳と心の研究推進室
室長　富田博秋（とみた　ひろあき）
電話番号：022-717-7262
E メール：htomita@med.tohoku.ac.jp

（報道担当）
東北大学東北メディカル・メガバンク機構
長神　風二（ながみ　ふうじ）
電話番号：022-717-7908
ファクス：022-717-7923
E メール：pr@megabank.tohoku.ac.jp

（AMED 事業に関すること）
日本医療研究開発機構（AMED）
疾患基礎研究事業部　疾患基礎研究課
電話番号：03-6870-2286
E メール：brain-pro＠amed.go.jp