子ども健康と環境に関する全国調査（エコチル調査）
論文概要の和文様式

雑誌における論文タイトル：

Prediction of gestational diabetes mellitus using machine learning from birth cohort data of the Japan Environment and Children's Study

和文タイトル：

エコチル調査のデータを用いた機械学習による妊娠糖尿病予測

ユニットセンター（UC）等名： 千葉ユニットセンター
サブユニットセンター（SUC）名：

発表雑誌名：Scientific Reports
年： 2023 DOI：10.1038/s41598-023-44313-1

筆頭著者名： 渡邉 応宏
所属 UC 名： 千葉ユニットセンター

目的：

近年、機械学習を用いて妊娠糖尿病の発症を予測する研究が報告されているが、生活環境の情報を多く含むデータの解析は行われていない。本研究では、エコチル調査のデータを用いて、各機械学習手法の性能とデータへの適応性の比較検討を行うとともに、妊娠糖尿病の発症に寄与する因子を網羅的に探索することを目的とした。

方法：

エコチル調査参加者のうち、妊娠糖尿病の既往歴のある 624 名と既往歴のない 82,074 名を対象とした。妊娠中の質問票調査、血液検査値、父親・母親の出生時体重等のデータを用いて、機械学習を用いた妊娠糖尿病の発症予測を行った。機械学習手法として、ランダムフォレスト（RF）、勾配ブースティング木（GBDT）、サポートベクターマシン（SVM）を使用し、比較対象としてロジスティック回帰（LR）を用いた。

結果：

各機械学習手法を比較したところ、GBDT、LR、RF、SVM の順で精度が良かった。GBDT モデルの「ROC 曲線下部面積」（疾患有無をどれだけうまく区別しているかを示す指標）は、妊娠糖尿病既往歴のある母親で 0.67（95%CI、0.59–0.75）、妊娠糖尿病歴のない母親で 0.76（95%CI、0.74–0.78）であった。探索的解析を行った結果、すでに報告されている因子に加えて、これまでほとんど報告されていない妊娠初期の主観的健康観や母親の出生時体重が妊娠糖尿病に影響する因子として検出された。

考察（研究の限界を含める）：

適切なアルゴリズムを使用することにより、エコチル調査のデータにおいても機械学習の手法を用いた探索的な解析を行うことができた。また、変数重要度の高かった変数は、先行研究で妊娠糖尿病との関連が報告されている変数が多く、予測モデルの妥当性が示された。その一方で、多様な情報を解析することにより、これまでほとんど報告されていなかった因子を検出することができた。本研究の限界として、利用可能なエコチル調査のデータセットのみで予測を行っているため、今回考慮できなかった他の因子、特に遺伝情報や糖尿病の家族歴の影響を含めた解析を行うことができず、そのため、予測精度自体は高くなかったことがあげられる。

結論：

決定木をベースとする GBDT、RF のアルゴリズムを用いた解析は、エコチル調査のデータセットに対して良好な精度を示し、より優れた学習モデルであった。妊娠糖尿病の発症を予測する重要な因子として、すでに報告されている因子に加えて、妊娠初期の主観的な健康観や母親の出生時体重といった因子が検出された。