【公開日】2026 年 1 月 7 日

作成日 2022 年 11 月 18 日
（最終更新日 2025 年 12 月 10 日）

「情報公開文書」（Web ページ掲載用）

受付番号：2025-4-142

課題名：機械学習手法の応用による疑似データ作成に関する研究

１．研究の対象

東北メディカル・メガバンク計画の地域住民コホート（宮城、岩手）・三世代コホートの参加者

２．研究期間

2022 年 12 月（研究実施許可後）～2031 年 3 月

３．研究目的

ゲノム医療推進に向けた最新のデータサイエンス関連人材育成には、質の良いデータが沢山必要になります。しかし、ゲノムデータや健康調査のデータは個人情報であり、その利用は慎重に行う必要があります。そこでこの研究では、データサイエンス人材が実データに近いデータで解析について学びをえることができるような疑似データの作成が可能かどうかを検討します。もっとも単純なアプローチとしては、健康調査票情報であれば、項目毎に実際の個人に紐付くデータの分布を調べ、その分布に従ったランダムな値を生成することで、項目毎のデータの集合として疑似データを作る事は可能ですが、この単純な手法で作成した疑似データは項目間の相関や遺伝情報と健康調査の項目の間の関係を持たないため、この疑似データを解析して得られる結果は非現実的な結果となり学習効果は限定的だと思われます。そこで本研究では近年研究が進む AI 系の手法を用いて、ゲノムワイド関連解析（GWAS 解析）や項目間の関連解析などのゲノム医療推進に標準的な解析に利用することができながら、参加者の方とは紐付かない擬似的なデータの作成を実際に行い、様々な解析によりその疑似データがどの程度ぐらい実データに近い形で利用できるかを検討します。

４．研究方法

ゲノムデータと健康調査票のデータをそれぞれ異なる方法で疑似データとし、その紐付けは元のゲノムデータから計算される検査項目と遺伝子変異の関連性の強さを評価するスコア（多遺伝子リスクスコアと呼びます）を使って行います。具体的には、ゲノムデータに関しては、メンデルの法則というよく知られた遺伝形式の法則にしたがって、ランダムに選んだペアから擬似的な子孫を生成します。これを 10 世代程度続けることで、現在の

データとの関連性が極めて低い擬似的なゲノムデータの生成を行います。調査票データに関しては、まず元のゲノムデータから計算した多遺伝子リスクスコアと元のゲノムに対応する調査票の項目を使って、多因子リスクスコアから擬似的な表現型データを作成するAIの開発を行います。次に、先に作成した疑似ゲノムデータから計算した多因子リスクスコアをこのAIに入力することで、対応する疑似表現型のデータを得る計画です。作成された疑似データが利用した元データと異なるデータであることは、値そのものが異なることや相関が低くなっているかどうかなど、多角的に検討を行います。
なお、本研究ではAI手法を用いた疑似データの作成とその結果の評価を行うのみで、個別の参加者の方の表現型と遺伝型の関連などの解析は実施しません。そのため、成果の論文発表や知財化を除いては、個別に開示するような結果はありません。

５．研究に用いる試料・情報の種類

全ゲノム情報、アレイ情報、生化学検査値、生理機能検査値、メタボローム解析値、及び調査票情報を用います。本研究では解析済みのデータのみを用い、試料を新たには利用しません。そのため、現時点で予測される不利益はありません。

６．外部への試料・情報の提供

解析に用いた元のデータの提供はありませんが、研究成果は学会や論文等で発表が予定されています。

7．研究組織

東北大学 東北メディカル・メガバンク機構　副機構長　木下賢吾

8．利益相反（企業等との利害関係）について

本学では、研究責任者のグループが公正性を保つことを目的に、情報公開文書において企業等との利害関係の開示を行っています。
本研究は、日本医療研究開発機構（AMED）の生命科学・創薬研究支援基盤事業研究費を使用し実施します。
本研究は、研究責任者のグループにより公正に行われます。本研究における企業等との利害関係については、現在のところありません。今後生じた場合には、東北大学利益相反マネジメント委員会の承認を得たうえで研究を継続し、本研究の企業等との利害関係についての公正性を保ちます。

9．お問い合わせ先

本研究に関するご質問等がありましたら下記の連絡先までお問い合わせ下さい。
ご希望があれば、他の研究対象者の個人情報及び知的財産の保護に支障がない範囲内で、研究計画書及び関連資料を閲覧することが出来ますのでお申出下さい。

照会先および研究への利用を拒否する場合の連絡先：

東北大学　東北メディカル・メガバンク機構　ゲノムプラットフォーム連携センター

TEL:　022-274-6040

研究責任者：

東北大学　東北メディカル・メガバンク機構　木下賢吾

◆個人情報の利用目的の通知に関する問い合わせ先

保有個人情報の利用目的の通知に関するお問い合わせ先：「９．お問い合わせ先」

※注意事項

以下に該当する場合にはお応えできないことがあります。

＜人を対象とする生命科学・医学系研究に関する倫理指針 第９章第20の1(3)＞

①利用目的を容易に知り得る状態に置くこと又は請求者に対して通知することにより、研究対象者等又は第三者の生命、身体、財産その他の権利利益を害するおそれがある場合

②利用目的を容易に知り得る状態に置くこと又は請求者に対して通知することにより、当該研究機関の権利又は正当な利益を害するおそれがある場合

◆個人情報の開示等に関する手続

本学が保有する個人情報のうち、本人の情報について、開示、訂正及び利用停止を請求することができます。

保有個人情報とは、本学の役員又は職員が職務上作成し、又は取得した個人情報です。

保有する個人情報については、所定の請求用紙に必要事項を記入し情報公開室受付窓口に提出するか又は郵送願います。詳しくは請求手続きのホームページをご覧ください。

（※手数料が必要です。）

【東北大学情報公開室】

https://www.bureau.tohoku.ac.jp/kokai/disclosure/index.html

※注意事項

以下に該当する場合には全部若しくは一部についてお応えできないことがあります。

＜人を対象とする生命科学・医学系研究に関する倫理指針　第９章第20の2(1)＞

①研究対象者等又は第三者の生命、身体、財産その他の権利利益を害するおそれがある場合

②研究機関の研究業務の適正な実施に著しい支障を及ぼすおそれがある場合

③法令に違反することとなる場合

以下、過去に掲載を行っていた文書

「情報公開文書」（Web ページ掲載用）

受付番号： 2022-4-128

課題名：機械学習手法の応用による疑似データ作成に関する研究

１．研究の対象

東北メディカル・メガバンク計画の地域住民コホート（宮城、岩手）・三世代コホートの参加者

２．研究期間

2022 年 12 月（研究実施許可後）～2026 年 3 月

３．研究目的

ゲノム医療推進に向けた最新のデータサイエンス関連人材育成には、質の良いデータが沢山必要になります。しかし、ゲノムデータや健康調査のデータは個人情報であり、その利用は慎重に行う必要があります。そこでこの研究では、データサイエンス人材が実データに近いデータで解析について学びをえることができるような疑似データの作成が可能かどうかを検討します。もっとも単純なアプローチとしては、健康調査票情報であれば、項目毎に実際の個人に紐付くデータの分布を調べ、その分布に従ったランダムな値を生成することで、項目毎のデータの集合として疑似データを作る事は可能ですが、この単純な手法で作成した疑似データは項目間の相関や遺伝情報と健康調査の項目の間の関係を持たないため、この疑似データを解析して得られる結果は非現実的な結果となり学習効果は限定的だと思われます。そこで本研究では近年研究が進む AI 系の手法を用いて、ゲノムワイド関連解析（GWAS 解析）や項目間の関連解析などのゲノム医療推進に標準的な解析に利用することができながら、参加者の方とは紐付かない擬似的なデータの作成を実際に行い、様々な解析によりその疑似データがどの程度ぐらい実データに近い形で利用できるかを検討します。

４．研究方法

ゲノムデータと健康調査票のデータをそれぞれ異なる方法で疑似データとし、その紐付けは元のゲノムデータから計算される検査項目と遺伝子変異の関連性の強さを評価するスコア（多遺伝子リスクスコアと呼びます）を使って行います。具体的には、ゲノムデータに関しては、メンデルの法則というよく知られた遺伝形式の法則にしたがって、ランダムに選んだペアから擬似的な子孫を生成します。これを 10 世代程度続けることで、現在のデータとの関連性が極めて低い擬似的なゲノムデータの生成を行います。調査票データに関しては、まず元のゲノムデータから計算した多遺伝子リスクスコアと元のゲノムに対応

する調査票の項目を使って、多因子リスクスコアから擬似的な表現型データを作成するAIの開発を行います。次に、先に作成した疑似ゲノムデータから計算した多因子リスクスコアをこのAIに入力することで、対応する疑似表現型のデータを得る計画です。作成された疑似データが利用した元データと異なるデータであることは、値そのものが異なることや相関が低くなっているかどうかなど、多角的に検討を行います。
なお、本研究ではAI手法を用いた疑似データの作成とその結果の評価を行うのみで、個別の参加者の方の表現型と遺伝型の関連などの解析は実施しません。。そのため、成果の論文発表や知財化を除いては、個別に開示するような結果はありません。

５．研究に用いる試料・情報の種類

全ゲノム情報、アレイ情報、生化学検査値、生理機能検査値、メタボローム解析値、及び調査票情報を用います。本研究では解析済みのデータのみを用い、試料を新たには利用しません。そのため、現時点で予測される不利益はありません。

６．外部への試料・情報の提供

解析に用いた元のデータの提供はありませんが、研究成果は学会や論文等で発表が予定されています。

8. 研究組織

東北大学東北メディカル・メガバンク機構　副機構長　木下賢吾

9. 利益相反（企業等との利害関係）について

本学では、研究責任者のグループが公正性を保つことを目的に、情報公開文書において企業等との利害関係の開示を行っています。

本研究は、日本医療研究開発機構（AMED）の生命科学・創薬研究支援基盤事業研究費を使用し実施します。

本研究は、研究責任者のグループにより公正に行われます。本研究における企業等との利害関係については、現在のところありません。今後生じた場合には、東北大学利益相反マネジメント委員会の承認を得たうえで研究を継続し、本研究の企業等との利害関係についての公正性を保ちます。

10. お問い合わせ先

本研究に関するご質問等がありましたら下記の連絡先までお問い合わせ下さい。
ご希望があれば、他の研究対象者の個人情報及び知的財産の保護に支障がない範囲内で、研究計画書及び関連資料を閲覧することが出来ますのでお申出下さい。

照会先および研究への利用を拒否する場合の連絡先：

東北大学　東北メディカル・メガバンク機構　ゲノムプラットフォーム連携センター

TEL： 022-274-6040

研究責任者：

東北大学　東北メディカル・メガバンク機構　木下賢吾

◆個人情報の利用目的の通知に関する問い合わせ先

保有個人情報の利用目的の通知に関するお問い合わせ先：「９．お問い合わせ先」

※注意事項

以下に該当する場合にはお応えできないことがあります。

＜人を対象とする生命科学・医学系研究に関する倫理指針 第９章第20の1(3)＞

①利用目的を容易に知り得る状態に置くこと又は請求者に対して通知することにより、研究対象者等又は第三者の生命、身体、財産その他の権利利益を害するおそれがある場合

②利用目的を容易に知り得る状態に置くこと又は請求者に対して通知することにより、当該研究機関の権利又は正当な利益を害するおそれがある場合

◆個人情報の開示等に関する手続

本学が保有する個人情報のうち、本人の情報について、開示、訂正及び利用停止を請求することができます。

保有個人情報とは、本学の役員又は職員が職務上作成し、又は取得した個人情報です。

保有する個人情報については、所定の請求用紙に必要事項を記入し情報公開室受付窓口に提出するか又は郵送願います。詳しくは請求手続きのホームページをご覧ください。

（※手数料が必要です。）

【東北大学情報公開室】http://www.bureau.tohoku.ac.jp/kokai/disclosure/index.html

※注意事項

以下に該当する場合には全部若しくは一部についてお応えできないことがあります。

＜人を対象とする生命科学・医学系研究に関する倫理指針　第９章第20の2(1)＞

①研究対象者等又は第三者の生命、身体、財産その他の権利利益を害するおそれがある場合

②研究機関の研究業務の適正な実施に著しい支障を及ぼすおそれがある場合

③法令に違反することとなる場合