試料・情報利用研究計画書（概要）

審査委員会 受付番号	2022-1017	利用形態	共同研究			
研究題目	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			研究期間	2023年1月～2026年3月	
代表研究機関	東北メディカル・メガバンク機構		責任者 氏名・職	木下　賢吾		教授
分担研究機関	東北大学未踏アナリティクスセンター		責任者 氏名・職	中尾　光之		センター長
研究目的と意義	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。						
研究計画概要	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。						
利用試料・情報	対象：コホート調査参加者　全員 試料：なし 情報：基本情報、調査票情報、検体検査情報、生理機能検査情報、ゲノム情報（全ゲノム、SNPアレイ）、メタボローム情報						
期待される成果	ゲノム医療でのAI系の手法の利用促進に資することが期待されます。						
倫理審査等の経過	2023年1月　東北メディカル・メガバンク機構倫理委員会承認						
倫理面、セキュリティー面の配慮	人を対象とする生命科学・医学系研究の倫理指針のほか、別途締結する研究契約を遵守して実施します。 利用する試料・情報は、東北メディカル・メガバンク機構スーパーコンピュータ内で、限定された研究者のみがアクセス可能な環境で利用します。						
その他特記事項	AMED補助金（生命科学・創薬研究支援基盤事業研究費）						

（事務局使用欄）

※公開日　　令和5年3月2日

※岩手医科大学いわて東北メディカル・メガバンク事業に協力された方で、本研究に限って試料・情報の利用を希望されない方は、下記までご連絡下さい。

岩手医科大学いわて東北メディカル・メガバンク機構　019-651-5110（5508/5509）