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　６ 年　 ６ 月　 ６ 日現在

機関番号：１１３０１
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１０４６５

研究課題名（和文）コホート間連携における調査票データクリーニングの自動化

研究課題名（英文）Automation of questionnaire data cleaning in inter-cohort collaboration

研究代表者
牧野　悟士（Makino, Satoshi）

東北大学・東北メディカル・メガバンク機構・助教

研究者番号：３０４２３４０３

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）：健常人ゲノムコホート・バイオバンクの必要性・重要性はますます高まっており、次のステップでは、既存のコホート研究との連携を推進し、健康・医療ビッグデータをさらに大規模化することが必須である。大規模なコホート連携に向けて、調査票による生活習慣・環境曝露の測定法についても標準化されたデータ取得・信頼性確保の方法が必要となる。そこで、調査票の様式が完全に同一でなくてもデータクリーニング可能な手法の開発を行った。

研究成果の学術的意義や社会的意義
近隣アジア諸国では、30万人から100万人を目標としたゲノムを含む分子疫学コホート構築が進んでおり、現在では数万人といった規模のコホートでこれらに対当していくことができない。データクリーニングと、それらのデータ統合及び精度管理を自動化することによって、コホート間の連携が可能となり、国際的評価にも耐えうる、日本人の代表性と公益性の高い、国民の健康の維持・増進、がんなどの疾病予防にとって質の高い大規模分子疫学コホート研究の構築が実現できる。

研究成果の概要（英文）：The necessity and importance of a healthy human genome cohort biobank is increasing, and the next step is to promote linkage with existing cohort studies, and it is imperative that health and medical big data be further scaled up. Toward large-scale cohort linkage, standardized data acquisition and reliability assurance methods are needed for lifestyle and environmental exposure measurement methods using questionnaires. Therefore, we developed a method that enables data cleaning even if the survey forms are not completely identical.

研究分野：ゲノム科学

キーワード：コホート研究　データクリーニング　外れ値検出

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1) 東北メディカル・メガバンク事業では、15 万人の一般集団の検体やコホート情報を有し、申請者らは取得された調査票データ、生理学機能検査データを初めとした各種のデータについて、統計学・機械学習の技法を用いたエラー候補の検出を進めてきた。次世代医療の社会実装にむけての取組みが本格的に進むなか、健常人ゲノムコホート・バイオバンクの必要性・重要性はますます高まっており、次のステップでは、既存のコホート研究との連携を推進し、健康・医療ビッグデータをさらに大規模化することが必須である。大規模なコホート連携のためには、調査票による生活習慣・環境曝露の測定法についても標準化されたデータ取得・信頼性確保の方法が必要となる。しかし、膨大なデータを全て人力で確認し、調査票原本に戻って修正の必要性を調べることは事実上不可能である。

(2) 近隣アジア諸国では、30 万人から 100 万人を目標としたゲノムを含む分子疫学コホート構築が進んでおり、現在の 15 万人コホートでは、これらに対当していくのに十分な大きさとはいえない。日本各地で展開の進む、JPHC 研究や J-MICC 研究といった分子疫学コホート研究との統合により、国際的評価にも耐えうる、日本人の代表性と公益性の高い、国民の健康の維持・増進、がんなどの疾病予防にとって質の高い大規模分子疫学コホート研究の構築が実現できる。そのようななか、データクリーニングと、それらのデータ統合及び精度管理方法は、大規模コホート研究のみならず、その重要性が認識されているものの、世界的にコンセンサスを得られた手法は存在していない。海外の大規模コホートである UK BioBank においても、タッチスクリーンベースであるためデータ入力時のエラー発生率は低いと考えられるものの、単純なミスマッチやデータ形式の違いを検出しているのみである。

２．研究の目的

(1) 本研究では、大規模なデータクリーニングにおいて、1) 集団からの外れ値を検出する際に既知の情報を利用して主成分分析（PCA）を拡張した統計的モデルを使用する、2) 検出されたエラー候補をその性質に基づいて分類・処理する、の二つの手法をあわせ用いることによってデータクリーニングを自動化し、データ取得方法と精度の違いによる統合困難化の回避を目的とした。

(2) 調査票における回答パターンの違いをエラー検出に利用し、これまで事実上不可能であった調査票の経時的データや家族間のデータのクリーニングに関しても応用可能なものとする。

３．研究の方法

(1) Mixed Model を使用した個体および家族内相関の調整

まず、R の glmer 関数に Mixed Model を適用し、複数の診断値（Marginal residuals, Conditional residuals, BLUP）を計算した。

(2) 質問項目ごとの外れ値を検出

時間軸（または家族）ごとにデータの特徴は異なるという仮定のもとに、各時間軸（または各家族）で外れ値を検出した。データをグループに分け、それに対して全体のデータとの差分に相当する情報を得た（混合モデルを適用）。次に、得られた回帰式からの予測値と実測値の差分（Marginal residuals, Conditional residuals, BLUP）を計算し、エラーの回答を外れ値に変換した。

(3) 複数質問における外れ値を検出

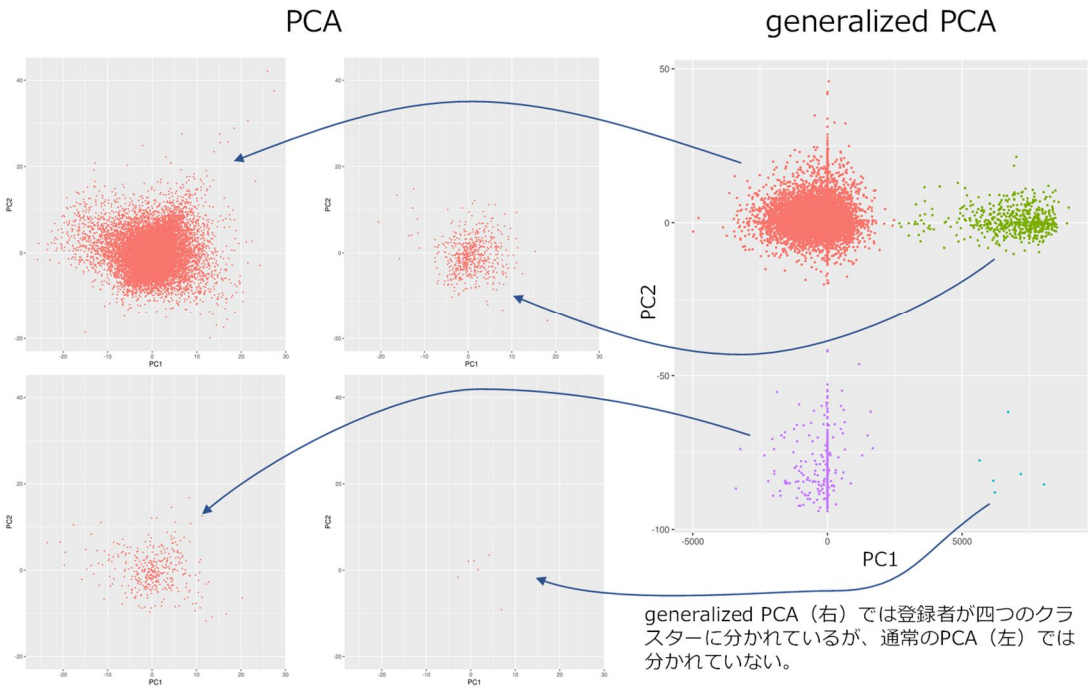
上記ステップで、エラーの回答を事前に外れ値に変換しておき、次に変換処理後の情報を使用して、さらに外れ値検出を行った。

４．研究成果

(1) 本研究では、集団からの外れ値を検出する際に既知の情報を利用して PCA を拡張した統計的モデルを使用し、調査票におけるエラー候補検出の自動化および精度向上のための手法を開発した。

(2) 質問項目ごとの外れ値検出結果（予測値と実測値の差分）を利用し、複数の質問をセットにした外れ値検出を、通常の PCA と generalized PCA で比較実施した。generalized PCA では、分布を仮定する手法のため 2 値データや離散データに特化したクラスタリングが可能となった。また、欠損値の補完が不要であり、重みを与えることができるといった利点がある。generalized PCA の結果では、通常の PCA ではみられなかったクラスター構造が検出されている（図）。

generalized PCAの結果で見られた4つのクラスター



generalized PCA（右）では登録者が四つのクラスターに分かれているが、通常のPCA（左）では分かれていない。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔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
研究分担者	田宮　元 (Tamiya Gen) (10317745)	国立研究開発法人理化学研究所・革新知能統合研究センター・チームリーダー (82401)	

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