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　３ 年　 ５ 月 ３１ 日現在

機関番号：１１３０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ１００９９

研究課題名（和文）大規模コホートの調査票における新規データクリーニング手法の開発

研究課題名（英文）Development of a new data cleaning method for questionnaires used in large cohorts

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）　　3,000,000円

研究成果の概要（和文）：大規模ゲノムコホート研究では、綿密な研究計画およびその計画に従った実施体制、エラー防止手法の導入にもかかわらず、種々のエラーの発生が不可避であり、それらのエラーは研究結果に大きく影響を与えるものとなりうる。しかし大規模な調査票をクリーニングすることは人力では不可能であった。そこで、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ことによって、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実装した。

研究成果の学術的意義や社会的意義
データクリーニングは、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なかった。海外の大規模コホートにおいても、多くは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。本研究はパターンの違いをエラー検出に利用するため、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となった。

研究成果の概要（英文）：In large-scale genomic cohort studies, in spite of careful study planning and implementation, and the introduction of error prevention methods, various errors are inevitable, and these errors may have a significant impact on the study results. However, it was not possible to manually clean a large number of questionnaires. Therefore, by using a statistical model that extends principal component analysis (PCA) by utilizing known information when detecting outliers from the data population, we developed and implemented a method to automate the detection of candidate errors and to improve its accuracy.

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

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

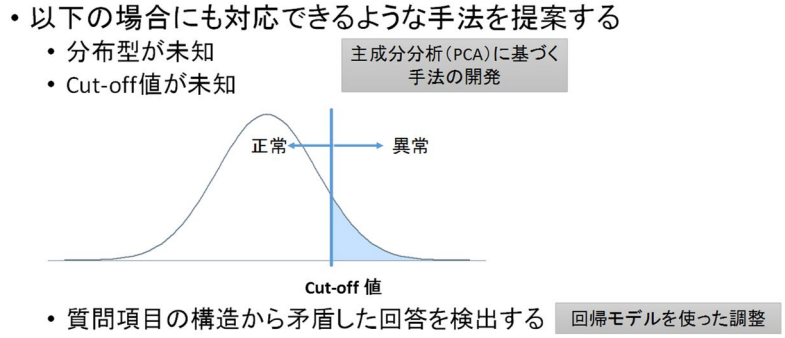
１．研究開始当初の背景

(1) 大規模ゲノムコホート研究では、綿密な研究計画およびその計画に従った実施体制、エラー防止手法の導入にも関わらず、種々のエラーの発生が不可避であり、それらのエラーは研究結果に大きく影響を与えるものとなりうる。調査の各段階で収集されるデータの精度を保ち、調査参加者への適切な調査結果の還元と医療支援への貢献、そして学術的用途としての有用性を高め、データの信頼性の確保に努めるためには、適切なデータクリーニングの実施が必須であり、薬剤疫学研究に関しては FDA などによる様々な「データの質」に関するガイドラインが存在しているが、データクリーニングの方法について標準となるものは、これまで明確には示されていない。特に、紙媒体の形態で収集される同意書、生活習慣などに関する調査票については、データ入力における手順の整備、モニタリングも必要となり、膨大な量となるデータを全て人力で確認すること、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った。また、論理的エラーは同一調査票内、同一登録者における調査票間（コホート調査への登録時のみならず、追跡調査など経時的に収集した調査票も存在する）、血縁関係にある登録者における調査票間（例えば、妊婦さんや生まれてくるお子さんを中心とした三世代コホートの場合）といった、非常に多岐にわたる関連を考慮しなければならない。

(2) 本研究は、紙媒体の形態で収集される情報の取扱いに主眼を置き、「どのような状態をエラーとするのか」「どのようにエラー候補の検出を行うのか」「検出されたエラー候補の取扱いをどうするのか」に関して、それぞれ新たなアプローチを行い大規模コホートにおける調査票データクリーニングにおける膨大なデータを取扱う上での問題の解決を目指すものである。

２．研究の目的

(1) 本研究は、同意書および調査票と、関連する特定健診データなどにおいて、期待した通りの挙動をしないデータ及びパターン（ここではエラーと呼称する）の検出を一つ目の目的とする。エラー検出に使われる手法は、19 世紀から集団からの外れ値を検出する統計モデルが存在し、近年ではニューラルネットワークやサポートベクトルマシン、決定木など分類問題を解く数理モデルを利用した手法が提案されている。後者は異常値の定義が明らかでない場合には適用できないが、前者は異常の定義が明確に数値化されていない状況でも検出可能であるという利点があり、今回対象としている医学研究のための一般集団コホートの参加者から得たデータのクリーニングに適している。なぜならば、この場合のエラーは、参加者の特異性の他に、参加者の回答ミス、データ入力者のミス、情報システム管理ミスによるデータの一貫性の欠落が原因で生じると考えたが、異常と正常とを分ける具体的な性質を予め知ることができないためである。ここで、エラーが一般コホート集団全体から見て、ごく少数の間で起きることを仮定することで、統計モデルが有用となる。また、エラーの分類として、我々は非現実的な値をとるデータだけでなく、論理的エラーが存在すると考える。エラー検出対象の登録データに対応する調査票には、一般集団コホートの中のある条件を満たす集団（例：女性、運動をすると回答した者、喫煙すると回答した者）のみに回答を求める質問項目が設計されており、回答対象に含まれない者からの回答が登録データに存在した場合にそれを論理的エラーと考える。このため、既知の情報として、我々は調査票の様式ファイル及び実際の登録データをあらかじめ読み合わせることにより、論理的エラーが生じうる箇所をリストアップすることが可能である。したがって、我々は統計モデルを用いてエラーの定義が明らかでない状況を許容するとともに、さらに既知の情報を取り込み、検出の精度を向上させる手法の開発および実装を行う。



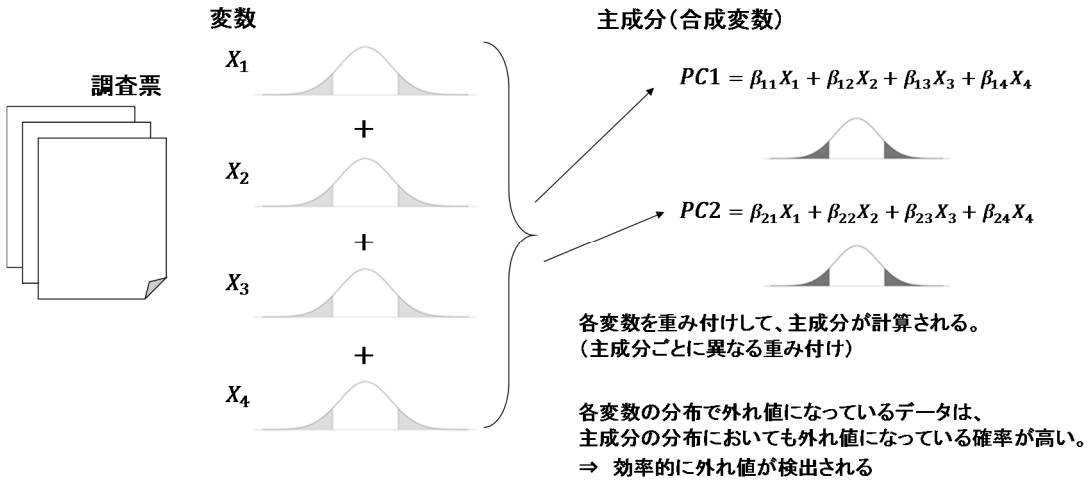
(2) 我々が開発する検出方法を適用して得られる結果をエラー候補とし、目視検査を通して精度の保証を図る。精度の保証が得られた段階で論理的エラーのリストアップの自動化の検討も行う。リストアップは、調査票の様式ファイルの読み合わせと、実際の登録データからの読みあわせを独立して実施する。目視検査作業は複数回実施し、エラー候補発見率の分布から、高頻度にエラーが生じうる箇所をほぼ全てリストアップしていることを統計的に推定できる。

(3) また、検出されたエラー候補の延べ数が非常に多数であり、調査票の原本に戻った修正では非現実的な時間と手間がかかることから、処理に工夫を行う。そのための調査として、エラー候補の性質（大まかに以下の二つと考えられる）に関する見積もりを行い、その結果に基づいて対処（合理的な情報としてそのまま分析する、欠測として扱う、単純なインピュテーション（平均値置換や最頻値置換）を施す、など）を決定する。
・原本の記述と乖離しており、原本に戻れば修正できる
・原本の記述と乖離せず、原本に戻っても修正できない

３．研究の方法

(1) 本研究におけるエラー候補検出に関する提案法の新規性は下記の 2 点である。
①人力による異常値検出から自動化アルゴリズムおよびソフトウェア実装
②既知の情報を取り込む工夫

①に関して、主成分分析に基づいた手法を開発する。主成分分析を用いた異常検出法はすでに多く提案されているが、さらに異常検出の自動化を行い、検出の作業効率を上げるため、統計量の一つである尖度を用いたアルゴリズムを開発する。また、開発したアルゴリズムはソフトウェアに実装し、実際の大規模コホートデータの解析に応用する。
②は、欠損パターンが既知である項目（例えば、女性にのみ回答を求める項目において、男性の試験参加者のデータは欠損である）に対し、想定された欠損パターンの情報をあらかじめモデルに入れておくことで異常パターンから除外し、検出結果の精度を向上させる。



さらに、検出された論理的エラー候補の性質応じた対処として、以下のような分類を行う。
・論理的矛盾に拘わらず、参加者の合理的な情報としてそのまま分析する。
・少数の参加者にしか見られないエラーを欠測として扱う。あるいは、単純なインピュテーション作業（平均値置換や最頻値置換）を施す。
・多数の参加者に見られるエラーは何らかの基準を設けることで合理的に対処する。
・これらの対処が可能にならない場合には、その項目全体を除去する。
・あるいは、エラー情報そのものを渡して特別な対処をしない。

４．研究成果

(1) 同意書および調査票と、関連する特定健診データなどにおいて、期待した通りの挙動をしないデータ及びパターン（＝エラー）の検出を一つ目の目的とした。そのために、これまで人力により行っていたエラーの検出を、主成分分析に基づいて自動化するようにアルゴリズムの開発を行った。エラー検出の自動化と、さらに検出の作業効率を上げるため、統計量の一つである尖度を用いたアルゴリズムとした。

(2) 開発した検出方法を適用して得られた結果をエラー候補とし、目視検査を通して精度の保証を図った。精度の保証が得られた段階で論理的エラーのリストアップの自動化の検討もあわせて実施した。リストアップは、調査票の様式ファイルの読み合わせと、実際の登録データからの読みあわせを独立して実施した。目視検査作業は複数回実施し、エラー候補発見率の分布から、高頻度にエラーが生じうる箇所をほぼ全てリストアップしていることを統計的に推定した。

(3) 開発した手法である、データ欠損のある変数のパターンを調整した上で(RAMP 法)、主成分分析によって主成分軸上での尖度を用いて外れ値を検出する機械学習手法(kurPCA 法)を、東北メディカル・メガバンク 15 万人分のデータに適用した。

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Narita A、Nagai M、Mizuno S、Ogishima S、Tamiya G、Ueki M、Sakurai R、Makino S、Obara T、Ishikuro M、Yamanaka C、Matsubara H、Kuniyoshi Y、Murakami K、Ueno F、Noda A、Kobayashi T、Kobayashi M、Usuzaki T、Ohseto H、Hozawa A、Kikuya M、Metoki H、Kure S、Kuriyama S	４．巻 10
２．論文標題 Clustering by phenotype and genome-wide association study in autism	５．発行年 2020年
３．雑誌名 Translational Psychiatry	６．最初と最後の頁 290
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41398-020-00951-x	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Okuda Hiroshi、Okamoto Koji、Abe Michiaki、Ishizawa Kota、Makino Satoshi、Tanabe Osamu、Sugawara Junichi、Hozawa Atsushi、Tanno Kozo、Sasaki Makoto、Tamiya Gen、Yamamoto Masayuki、Ito Sadayoshi、Ishii Tadashi	４．巻 24
２．論文標題 Genome-wide association study identifies new loci for albuminuria in the Japanese population	５．発行年 2020年
３．雑誌名 Clinical and Experimental Nephrology	６．最初と最後の頁 1～9
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10157-020-01884-x	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Tadaka S, Katsuoka F, Ueki M, Kojima K, Makino S, Saito S, Otsuki A, Gocho C, Sakurai-Yageta M, Danjoh I, Motoike IN, Yamaguchi-Kabata Y, Shirota M, Koshiba S, Nagasaki M, Minegishi N, Hozawa A, Kuriyama S, Shimizu A, Yasuda J, Fuse N; Tohoku Medical Megabank Project Study Group, Tamiya G, Yamamoto M, Kinoshita K.	４．巻 6
２．論文標題 3.5KJPNv2: an allele frequency panel of 3552 Japanese individuals including the X chromosome	５．発行年 2019年
３．雑誌名 Human Genome Variation	６．最初と最後の頁 28
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41439-019-0059-5	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Sakurai Rieko、Ueki Masao、Makino Satoshi、Hozawa Atsushi、Kuriyama Shinichi、Takai-Igarashi Takako、Kinoshita Kengo、Yamamoto Masayuki、Tamiya Gen	４．巻 印刷中
２．論文標題 Outlier detection for questionnaire data in biobanks	５．発行年 2019年
３．雑誌名 International Journal of Epidemiology	６．最初と最後の頁 1305～1315
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/ije/dyz012	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Sakurai Rieko、Hattori Satoshi	4．巻 2
2．論文標題 Goodness-of-fit test for the parametric proportional hazard regression model with interval-censored data	5．発行年 2018年
3．雑誌名 Biostatistics & Epidemiology	6．最初と最後の頁 115～131
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/24709360.2018.1529347	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 櫻井利恵子
2．発表標題 バイオバンクにおける質問票データに対する外れ値検出
3．学会等名 第2回日本メディカルAI学会学術集会
4．発表年 2020年

〔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)	東北大学・東北メディカル・メガバンク機構・教授 (11301)	
研究分担者	櫻井　利恵子 (SAKURAI Rieko) (50794541)	東北大学・東北メディカル・メガバンク機構・非常勤講師 (11301)	

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