厚生労働科学研究費補助金（がん対策推進総合研究事業）
分担研究報告書
匿名化手法の検討・評価に関する研究

研究分担者　南　和宏　統計数理研究所教授
研究分担者　東　尚弘　国立がん研究センターがん登録センターセンター長
研究分担者　柴田　亜希子　国立がん研究センターがん登録センター全国がん登録室室長

研究要旨: がん登録情報の安全な学術研究利用には、匿名化手法の確立が不可欠である。本研究では、代表的な匿名化手法である *k*-匿名化の安全性評価に必要となる漏洩シナリオ、それに対応する攻撃者モデルを定式化し、セキュリティ・パラメータの決定に必要な最小セル度数に関する評価実験を行った。その結果、詳細な地域情報、疾病分類コードによるレコード識別リスクは高く、さらに単独では識別リスクが高くない性別、年齢等の基本的属性情報を複数組み合わせることで、識別リスクが高まることが判明した。

A．研究目的

全国がん登録の情報には、医療機関の受診者に関する機密情報が含まれており、がん登録情報を用いた調査研究を行う際に、匿名化データからの機密情報が外部に漏洩しないような安全性の担保が必要である。現在、匿名化データの代表的な安全性指標として、*k*-匿名性および、その派生指標が多く提案されている。しかしがん登録情報に対して具体的にどの手法を選択すべきかその要件は明らかでない。本研究では、匿名化手法の安全性の評価手法の確立を目指し、情報漏洩に関する攻撃者モデルの定式化を行い、外観識別性の高い属性情報によるレコード識別リスクの評価を行った。

B．研究方法

本年度は、代表的な安全性指標である *k*-匿名化の枠組みで匿名化処理の問題定式化を行った。*k*-匿名化では、レコードの属性情報を外観識別性の高い準識別子と機密属性に分類することが要件になるため、その判断の前提となる現実的な攻撃者モデルを３種類定義し、それぞれのモデルにおける属性情報の外観識別性を３段階で評価した。また実際の匿名加工処理では情報の粒度を制御する必要があるため、各準識別子情報について、情報粒度の関係性を記述すドメイン一般化階層を定義した。

さらに実際のがん登録情報を用い、具体的な *k*-匿名化の実施アルゴリズム、セキュリティ・パラメータ（最小セル度数）の決定に必要なレコード識別リスクの評価実験を行った。評価実験では外観識別性の高い代表的な属性を単一、または複数の組み合わせでクロス集計した場合の度数の頻度分布を算出し、識別リスクの高い属性の種類、値の範囲を明らかにした。

C．研究結果

単一属性としては、「診断時年齢」、「市区町村コード」、「ICD10 コード」等を用い、個々の値の頻度分布を解析した。その結果年齢については、108 歳を超える高齢の受診者のレコード数は 10 未満と識別リスクが高く、ある年齢以上の情報をグループ化するトップコーディングの処理が必要であることが分かった。また市区町村コード、ICD10 についても度数が 10 以下の市区町村コードが 50 以上存在し、これらの値はカテゴリー変数であるため匿名化処理における適切なグループ化方法が重要な検討事項であることを確認した。

複数属性の組み合わせとして、基本的属性である「性別」、「年齢」、「都府県コード」の組み合わせを組み合わせたクロス集計における頻度分布を解析した。これら３つの属性のクロス分析の結果、人口の少ない都道府県では、20 代の若年層、90 代後半の高齢者で頻度 10 未満のセルが多く存在する

ことが明らかになった。このように単独では識別リスクが低い属性であっても複数属性の値による絞り込みで識別リスクが高まることが分かり、多次元データの適切なクラスタリングが、匿名化処理における今後の重要な検討事項であることを確認した。

D．考察

がん登録情報に対するレコード識別リスクを評価し、地域情報、病名コードに個人を特定する可能性の高い稀な値が多く含まれることが分かった。よってこれらの情報は匿名化処理における一般化処理の対象となるが、地域情報、病名に関するドメイン一般化階層を決めるにあたり、どのように情報の粒度、階層構造を決めるべきか明らかでない。今後は、匿名データの有用性とドメイン一般化階層の性質の関係性を分析し、一般化処理に関する設計原理を探求する予定である。

また単独の情報としては識別リスクが低くとも複数の組み合わせにより、識別リスクが著しく高まることも今回の評価実験で明らかとなった。したがって匿名化処理を行う際、個々の属性情報ごとに一般化処理を行うのではなく、多次元データのデータ空間の領域分割問題として匿名化の問題を捉え、柔軟なデータのクラスタリングを行う匿名化アルゴリズムが必要と言える。既存研究でもそのような方向性の匿名化アルゴリズムがいくつか提案されているが、データ空間の分割方法が単純なものが多く、匿名化データの安全性と有用性（情報損失）の有用性とトレードオフが適切に取られるかについては実証的に評価する必要がある。また *k*-匿名化処理は NP 困難と分類される最適化問題の一つに分類されるため、多次元データを対象とした場合の計算効率の評価を行い、必要があれば近似的なアルゴリズムを検討する予定である。

E. 結論

がん登録情報のレコード識別リスクを実証的に評価し、単独または複数の属性情報による識別リスクの高くなる条件を明らかにした。今後はそのような高リスクの状況を柔軟に回避する匿名化処理手法の開発に取り組む予定である。

F. 研究発表

1.　論文発表

特になし。

2.　学会発表

1.南和宏. 令和 2 年度革新的自殺研究推進プログラム 第 1 回ワークショップ. 公的ミクロデータのプライバシー保護と表データの秘匿処理ツールの紹介. 2021 年 2 月 15 日.

2. 小野　元，南 和 宏．Toward Locally Private Logistic Regression with Missing Data. コンピュータセキュリテイシンポジウム 2020. 2020 年 10 月 26 日.

3.南和宏. 差分プライバシーと匿名化. コンピュータセキュリテイシンポジウム 2020. 2020 年 10 月 28 日.

G. 知的財産権の出願・登録状況

特になし。