別紙３

厚生労働科学研究費補助金（がん対策推進総合研究事業研究事業）
研究報告書

全国がん登録の円滑な運用に向けた研究

研究代表者　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

研究要旨
がん登録等の推進に関する法律に基づき全国がん登録は 2016 年診断症例以降、全国の病院から義務的届出が開始され、2019 年に初年罹患数が始めて発表され 2020 年には 2017 年罹患数が発表されたが、それらの数の動きは制度の変わり目による影響が考えられる。この制度変革による罹患数の影響は今後も追跡して安定を検討すべきである。また、他にも予後情報の精度については検証すべき、また、利活用における提供データの安全性についても定量的な評価による検討が行われているべきであると考えられる。本研究においてはこれらの検討を目標に、それぞれ解析を計画し、初年度においては、その準備となるような検討・解析を行った。具体的には、制度安定化指標としての初診届出不明数／割合の検討、予後情報の住民票照会に基づく検討、また、k-匿名化達成の解析、を行った。今後これらの解析を進めることで、円滑な全国がん登録システムの運営を目指す。

研究分担者氏名・所属研究機関名・職名

東　尚弘　　国立がん研究センター
がん対策情報センター
がん登録センター　センター長

祖父江　友孝　大阪大学
大学院医学系研究科
教授

柴田　亜希子　国立がん研究センター
がん対策情報センター
がん登録センター
全国がん登録分析室長

南　和宏　　統計数理研究所
モデリング研究系
准教授

A. 研究目的

がん登録等の推進に関する法律に基づき全国がん登録は 2016 年診断症例以降、全国の病院から義務的届出が開始され、2019 年に初年罹患数が 995,131 例発表された。これは前年の 2015 年地域がん登録の罹患数 903,914 例から約 9 万例の増加であり、地域がん登録の毎年数万例程度の増加に比べると急な増加である。これは制度移行の影響と考えられている。

全国がん登録制度の運用の安定化と改善と信頼のためには①データの質評価が必要不可欠である。さらに、未着手の課題として、2019 年度から始まった全国がん登録情報の提供の②データ匿名化の安全性評価の確立の 2 点が必要である。以上のように、本研究は特にデータの質と安全なデータ利用について、今後の全国がん登録制度の健全な運営を確保するための上記検証活動を行い今後の体制に反映させることが目的である。

B. 研究方法

① データの質評価

細分化すると a.登録数や情報内容の質、及び、b.死亡情報の突合確率、の二つが要検討である。a 登録数については前述の制度移行の影響が、届出件数、治療開始後の届出割合、既登録との突合確率、遡り調査回答の診断年分布などの 処理過程の各段階における症例数を記述し観察することで影響の大きさを検討し、適切な指標を同定する。

b.死亡情報については、これまで国からの死亡情報を提供されている院内がん登録や一部の地域がん登録で行われていた住民票照会による生存状況確認との差異が生じる可能性がある。そこで従来の

住民票照会を 2016 年症例サンプルについて 1、3 年目に行う。

②データ匿名化の安全性評価の確立

細分化すると a. 匿名化個票の提供における安全性確保、b.データ公表における秘匿性と有用性確保のバランスの 2 種類の焦点がある。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、各準識別子情報について、情報粒度の関係性を記述する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。

C. 研究結果

本年度は、3 年計画の 1 年目ということで、以下を行った。

① データの質評価

・登録数や情報内容の質は、制度としての安定性に関連していることから、その制度安定性の指標を検討した。制度移行における罹患統計への影響を反映した指標としては、初診届出不明例の数、割合が考えられた。また、前届出件数、整理症例数割合なども指標として考えられた。今後、実際の算出も検討する。

・予後情報の精度を検討するために国立が研究センター中央病院の 2016 年症例の通院継続者を除く症例に対しての住民票照会による追跡を行った。3,824 人を調査、3,749 人に関しての住民票照会が可能であり、死亡 2,343 名、生存 1,333 名、不明 73 名（追跡不能 72 名、除票 1 名）であった。この結果を 3 年目に全国がん登録と情報と突合して検討する。

② データ匿名化の安全性評価の確立

全国がん登録匿名データの申出を行い、匿名化データの安全性の基準として、手始めに、一般的に入手可能な情報を要素として k-匿名化の評価を行った。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。年齢については、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」、「都道府県コード」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、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

本年は、3 年計画の 1 年目ということで 2 年目以降のための集計方法（指標）の検討、準備データの収集、データ解析を開始した。全国がん登録の円滑な運用のために必要な要素についての準備的な解析を進めた。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）とで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

全国がん登録の制度は開始後 5 年が経とうとしているが、そのデータ活用は始まったばかりである。今年度は、データの質評価及びデータ匿名化の安全性評価のための準備データの収集、データ解析を開始したが、円滑な運用のために試行錯誤しながら、本研究班においてその検討を深めていく。

F. 研究発表

なし

G. 知的財産権の出願・登録状況

なし