厚生労働行政推進調査事業費補助金（がん対策推進総合研究事業研究事業）
（総合）研究報告書

全国がん登録の円滑な運用のための検証に関する研究

研究代表者　　東　尚弘　　国立がん研究センター がん対策研究所 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 年に初年罹患数が初めて発表され、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、予後情報の精度や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るべきであると考えられる。
本研究では、これらの検討を目標に、それぞれ解析を計画し、最終年度では、データの質評価として制度移行の影響のモニタリング、予後情報の精度について評価を行った。また、データ匿名化の安全性評価の確立を目指し、提供されるデータの安全性について、k-匿名化による評価・検討を行ったほか、医療系ミクロデータであるがん登録情報を対象に、地域情報の匿名化を柔軟に行う匿名化アルゴリズムを開発し、その有用性に関する実証的評価を行った。
これらの研究結果から、診断施設不明例は、制度安定化を評価するための指標の一つになると考えられた。また、k-匿名化及び匿名化アルゴリズムの活用により、より安全かつ有用な全国がん登録情報の提供が可能になると考えられた。

研究分担者氏名・所属研究機関名・職名

研究代表者　　東　尚弘　国立がん研究センター がん対策研究所 がん登録センター センター長
研究分担者　　祖父江　友孝　大阪大学・大学院医学系研究科・教授
研究分担者　　増田　昌人　　琉球大学病院 がんセンター　センター長　診療教授
研究分担者　　南　和宏　　統計数理研究所 データ科学研究系　教授
研究分担者　　塚田　庸一郎　国立がん研究センターがん対策研究所がん登録センター　院内がん登録室長
研究分担者　　柴田　亜希子　国立がん研究センターがん対策研究所がん登録センター 全国がん登録分析室長（令和 4 年度まで）
研究分担者　　榊原　直喜　　国立がん研究センターがん対策研究所がん登録センター　全国がん登録分析室　研究員
研究協力者　　藤下　真奈美　国立がん研究センターがん対策研究所がん登録センター 全国がん登録室長

A．研究目的

がん登録等の推進に関する法律（がん登録推進法）に基づく全国がん登録では 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 年に初年罹患数が初めて発表され、がん罹患数の報告書が厚生労働省から公開されている。全国がん登録を真に役立て円滑に運用していくためには、まず第一に罹患数の正確性の確保が重要であり、第二に二次利用データの安全性を確保していく必要がある。

1 点目の罹患統計の質について検証するための動きはまだ無い。全国がん登録の質について懸念される点は大きく分けて 2 点、①罹患数が正しいか、②死亡情報が正しいか、が想定される。①については、これまでがん罹患数が 2016 年で急増、2017 年は漸減と、その動きはおそらく制度の変化による影響を色濃く反映していると考えられ、正

確な罹患数とは言えないことを示している。この原因として診断年の整理の不十分さが考えられており、今後の罹患統計が安定していくには一定の時間がかかると考えられるとともに、その安定化の進行をモニターする指標が必要と考えられる。伝統的・国際的には、DCO(Death Certificate Only)、DCI(Death Certificate Initiation)ががん登録の精度指標とされるが、この指標は診断年整理においては指標とならない。

②死亡情報の正確性については、わが国の全国がん登録が、氏名などの個人識別情報を使って死亡票とがん登録の既登録情報を連結させ、連結されなかったものは「生存」とされることから、その連結精度を検証することが必要である。特に個人識別情報が不十分なために連結されないものが生ずると、当該症例はデータベース上では永遠に生存したままになってしまうため、長期に予後を追跡するほど影響は大きくなる。そのため、この検証は非常に重要であると考えられる。

2 点目の二次利用データの安全性については、その提供可能性にかかわる課題である。主として、提供する個票データの安全性と、公表する集計データから個人が同定されないか、といった安全性の問題が存在するが、本研究においては、結果的には前者を中心に検討を進めることとなった。

本研究は以上のように、全国がん登録を円滑に運営していくための知見を提供することを目的としている。

B． 研究方法

①データの質評価

1) 届出数及び情報内容の質の評価

全国がん登録の届出数や情報内容の質を評価するため、全国がん登録の運営上で算出される指標を検討した。その中で診断年の整理に活用可能な指標を設定して、その後の制度移行の影響などについてのモニタリングを行った。

2) 予後情報の精度の評価

また、予後情報の精度を評価するために、国立がん研究センター中央病院の院内がん登録の2016年症例、2017年症例（通院継続者を除く）について、住民票照会による追跡等で生存状況の評価を行い、それらを国立がん研究センター中央病院の所在地である東京都からのがん登録推進法第20条に基づく届出症例に対する生存確認情報の返却データと比較した。

②データ匿名化の安全性評価の確立

提供における匿名化個票の安全性確保、データ公表における秘匿性と有用性確保のバランスについて以下のような検討を行った。

1) 匿名化された情報の提供における安全性の検討

2016 年、2017 年の匿名化された全国がん登録データを用い、提供されるデータの安全性について、k-匿名化による評価・検討を行った。

2) 全国がん登録情報の匿名化指標の開発

がん登録情報の地域情報に国土交通省の位置参照情報を結合し、地域の位置座標に基づき地域領域を柔軟に分割する匿名化アルゴリズムを開発した。この提案手法の有効性を示すため、従来の地域レベルの調整による匿名化アルゴリズムも合わせて実装し、匿名処理で生成されるグループ間の均一性を有用性の指標として両者の比較を実証的に行った。

C． 研究結果

①データの質評価

1) 制度移行の影響の評価

制度移行の影響のモニタリング指標として、「診断施設不明割合」を考案した。ある患者が年を越えて複数の病院を受診し、それぞれの施設から届出があっても名寄せが成功すれば統合され、また診断年月日が最初のものが、当該患者の診断年となる。それと同時に、その施設が「診断施設」となる。また、最初の施設からの届け出がない、あるいは名寄せが全部はまとまらないということが生じても、名寄せが成功したうちの最も早い日が診断年となる。しかし、症例区分により既に診断されていたものであれば、「診断施設」が不明となる。診断年は名寄せが成功しても変わらないことはあり得るが、診断年変更の割合がほぼ一定とした場合には、診断施設不明割合は、名寄せの成功率に比例すると考えられるため、診断年の確からしさを表す指標として活用可能と考えた。実際に、この値の経過を集計値で追跡したところ、2016 年は 69,141 例（7.0%）、2017 年は 59,606 例（6.1%）、2018 年は 54,489 例（5.6%）、2019 年は 49,482 例（5.0%）と漸減傾向であったこと、また、罹患数も以前の経過から大きく外れなくなってきたように見えることから、安定化の指標として活用できる

と考えられた。

2）死亡票の突合率

予後情報の精度については、2019 年の全国がん登録の死亡情報と突合したところ、院内で生存状況が確認できた 16,890 名（生存 11,327 名、死亡 5,563 名）のうち、死亡が確認されている者で、全国がん登録でも死亡が確認できた者は 5,529 例（99.4%）、確認できなかった者は 34 名（0.6%）であった。また、院内で生存が確認された者は全国がん登録でも全て生存が確認でき（100%）、ほぼ実態に近い生死状況を把握できていた。

②データ匿名化の安全性評価の確立

1）匿名化された情報の提供における安全性の検討

基本的な安全性確認のため、ICD-10 のみ、ICD-O-3 の部位コードのみ、ICD-O-3 部位コードと組織型コード、さらに性別、年齢を組み合わせた時の k-匿名化の評価として、ユニーク（k=1）となる症例を集計した。ユニークになるものは ICD-O-3 の部位分類のみで 58 件、ICD-10 分類では 86 件、部位組織分類まで含めると 4,639 件であった。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。

2）全国がん登録情報の匿名化指標の開発

今回の提案手法をがん登録情報の住所情報に適用したところ、既存の地域レベルを調整する匿名化アルゴリズムと比較して、グループ間の均一性を定量化する Discenability 指標において、5%から 16%の改善が確認された。また同一グループに含まれる地域情報の隣接性についても従来手法の結果に比べて際立った改善が実現できることが示された。

D． 考察

全国がん登録制度の安定化を図るには、データの質評価が重要である。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おり、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、診断施設不明例の数・割合、前届出件数、整理症例数割合などが考えられる。本研究では、モニタリング指標として診断施設不明例を用いたが、最新年では 5～6%の減少傾向を認めており、今後も精度は向上していくものと考えられた。

また、予後情報の精度については、国立がん研究センター中央病院の院内がん登録症例を用い、住民票照会による追跡等で評価を行った。登録精度については、都道府県によって多少ばらつきがあることに留意する必要がある。

提供されるデータの安全性についての k-匿名化による評価・検討では、ICD-O-3 の部位分類のみ、ICD-10 分類のみ、部位組織分類のみの場合においてユニーク（k=1）となる症例を削除しても、全体の件数は 200 万件以上のためデータの有用性という意味では特に問題ないと思われた。一方で、ICD-O-3 の部位・組織分類や ICD-10 分類、性別、年齢を加えるとユニークな症例が増えるため、必要な項目とその有用性に応じて検討をする必要があると考えられた。

E． 結論

データ提供における匿名化個票の安全性確保、データ公表における秘匿性と有用性確保のバランスの双方に関して、これまでの検討を踏まえた解析を行った。

これらの研究結果から、「診断施設不明例」は、制度安定化を評価するための指標の一つになると考えられた。また、k-匿名化及び匿名化アルゴリズムの活用により、より安全かつ有用な全国がん登録情報の提供が可能になると考えられた。

F．健康危険情報

特になし

G． 研究発表

特になし