# がん登録情報の匿名化された情報とプライバシーリスクについて

柴田 亜希子[1] 山岡 裕司[2] 小栗 秀暢[2] 石黒 利幸[3] 安藝 理彦[3] 吾妻 光洋[3]
[1]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
[2]株式会社富士通研究所セキュリティ研究所
[3]富士通株式会社第二ヘルスケアソリューション事業本部

## 1. はじめに

個人情報保護法上、個人情報ではないものの、取扱いに一定の規律が課される情報として匿名加工情報がある。匿名加工情報とは、個人情報を加工して、①特定の個人を識別できない、②元の個人情報を復元できない、ようにしたものをいう。「特定の個人を識別できる」とは、情報単体又は複数の情報を組み合わせて保存されているものから社会通念上そのように判断できるものをいい、一般人の判断力又は理解力をもって生存する具体的な人物と情報の間に同一性を認めるに至ることができること、とされている。匿名加工情報は、個人情報保護法の目的である情報に関する個人の権利利益を保護しつつ、情報の利活用を促進することを目的に定義された。守るべき個人の権利利益の実態には諸説あるが、プライバシー権の一種である「自己情報コントロール権」が有力である。本研究は、2019 年 1 月から提供開始予定の全国がん登録情報の匿名化された情報の提供を想定し、既存のリスク評価技術を用いて、がん登録という特殊な集団の個人データの持つプライバシーリスクを可視化して理解することを目的とした。

## 2. 方法

プライバシーリスクの評価には、既存の、確立された評価技術（株式会社富士通研究所）を用いた[1]。このプライバシーリスク評価技術は、個人情報の漏えい事件による損害賠償の複数の判例から設定された、個人情報価値の算定モデル「J0 モデル」（NPO 法人 JNSA）を、匿名加工情報向けに拡張したものである。J0 モデルは、個人情報が漏えいした場合の個人の精神的苦痛レベル（P レベル）と経済的損失レベル（E レベル）をそれぞれ大中小の 3 群に分けて情報機微度（EP レベル）として定量化し（図 1）、さらにその個人情報の個人の特定し易さを本人特定容易度として定量化し、それらを掛け合わせて金額としてプライバシー侵害リスクを表現する。

個人情報価値（円／人）=基礎情報価値×機微情報度×本人特定容易度

J0 モデルの本人特定容易度は、氏名や電話番号のない情報に対する算定値がどのような匿名加工をしても一定になってしまい、匿名加工情報のリスク評価には適さない。そのため、プライバシーリスク評価技術では、加工されたデータの分布や情報機微度に基づいて本人特定容易度を算定するよう拡張している。また、一般人が観察できない情報は「観察可能性なし」と設定することで本人特定容易度の算定の対象外にできる。よって、個々の情報の EP レベルと観察可能性の設定によって、評価結果が変わる。

図 1　がん登録情報の EP レベル

我々は、 2014 年の山形県（人口約 116 万人）のがん罹患（10,542 件）の研究目的提供情報について、EP レベルと観察可能性を設定し（図 1)、匿名加工の程度とプライバシーリスクの関係について、個人情報価値による評価を試みた。

匿名加工は､独立行政法人等の保有する個人情報の保護に関する法律についてのガイドラインの「非識別加工情報」の加工基準を参考として、近年、κ≧3 のκ-匿名性が匿名性の指標になりつつあることから、第 1 号の識別子である年齢と住所について κ≧3 の匿名化を図ることを目標に、情報の抽象化レベルを変えて、その他の情報がある場合を評価した。また､がん登録情報のうち経済的損失リスクが最も高い情報を仮にがん病名として､ICD-10 の EP レベルを 22 とした場合の個人情報価値の変動の程度も計測した。

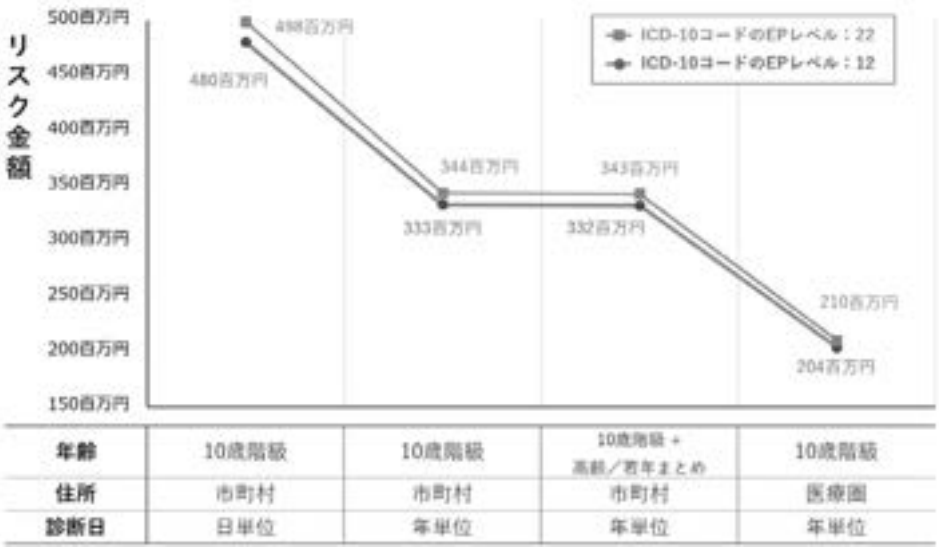


図 2　匿名加工の程度とプライバシーリスクの関係

3. 結果

約 1 万件の 1 年分のがん罹患データの場合、基本的な匿名加工である診断時住所の市区町村置き換え、年齢の 10 歳階級グループ化だけでは、 $\kappa \geqq 3$ の匿名化はまったく満たせないことが分かった（図 2)。一律に診断年に置き換えることにより、当初よりプライバシーリスクが約 31%下がることが分かった。年齢については、罹患数の少ない 90 歳代以上と 20 歳代以下をまとめても、10 歳年齢階級と比べてリスクの程度は同等であった。住所について、山形県の 35 市町村区分では特異な記述となりやすい市町村が出現することから、4 医療圏に置き換えると、年齢区分が 10 歳階級のままでもリスクがさらに下がった。

4. 考察

がん登録情報は腫瘍の局在と種類の組み合わせだけでも 10 万件以上のため、観察可能性ありの全属性に対して $\kappa \geqq 3$ の匿名性を満たすことが難しく、高度に匿名加工の技術を適用しない限り、高リスクに評価されることが分かった。また、がん登録情報は、まれな事象（外れ値）について基本的で効果的な匿名加工処理のうちの「削除」を行うと、関心の調査研究が成立しない特性を持つ。今回は機微情報度が高い属性を加工していないため、リスクは本人特定容易度に比例しており、ICD-10 コード（病名）の E レベルの設定は個人情報価値に大きく反映しないことが分かった。さらに、この評価技術では、側性の「両側」のように、特定の局在にのみ用いる区分は、全データ中では希な頻度のため、直感的な個人情報価値よりも高リスクに評価されることが分かった。また、EP レベルは、属性単体の性質で決めるため、複数の属性の組み合わせによって P レベルが高くなるケース（男性の卵巣癌等）でも、個人情報価値が正しく反映されない場合があることが分かった。

5. 結論

既存のプライバシーリスク評価技術を用いて、がん登録のように特殊な集団の匿名化された情報でも、一般人の判断力や理解力による個人識別リスクについて、一定の判断基準を提供するだろうことを確認できた。がん登録に従事する者、調査研究に利用する者は、がん登録の情報は、本人への説明なく収集され、同意なく第三者に提供できる情報であることから、国民の信頼を裏切ることなく、プライバシーの保護と当該情報を用いた調査研究がもたらす結果の公益性のバランスに十分配慮して「匿名化された情報」を取り扱う必要がある。

---

引用文献

[1] 山岡裕司，伊藤孝一．k-匿名性による特定可能性分析に基づいたデータプライバシーのリスク分析．情報処理学会研究報告．2016：1-8.