投稿

院内がん登録における匿名化手法の検討

渡邊 多永子（ワタナベ タエコ）＊1　東 尚弘（ヒガシ タカヒロ）＊2　山城 勝重（ヤマシロ カツシゲ）＊3　海崎 泰治（カイザキ ヤスハル）＊4　津熊 秀明（ツクマ ヒデアキ）＊5
固武 健二郎（コタケ ケンジロウ）＊6　猿木 信裕（サルキ ノブヒロ）＊7　岡村 信一（オカムラ シンイチ）＊8　柴田 亜希子（シバタ アキコ）＊9　西本 寛（ニシモト ヒロシ）＊10

目的　全国で指定されているがん診療連携拠点病院から提供される院内がん登録全国データは，わが国のがん診療の現状を示す貴重なデータである。このデータの個人情報保護特性を明らかにし，より多くの研究者が解析利用できるように，匿名化手法の検討を行った。

方法　まず，院内がん登録データ2008年症例（N＝424,983）のリスク評価を行った。次いで，全データで一律にキー変数の情報量を減らす加工（大域的再符号化）とリスクの再評価を行い，さらに安全性を上げるためにリサンプリングの効果も検討した。リスク評価には「一意」の数・割合を用いた。

結果　病院名削除と年齢の処理で，一意割合は1.7～3.0％にまで低下したが，母集団が大きいことから依然として7,140～12,699が一意であった。ランダムサンプリングを行うと，例えば50％の抽出率で，抽出後の標本で一意に見えるデータの約半数が母集団一意でない状態となった。

結論　診断病院名の削除と年齢のグルーピングやトップ／ボトム・コーディングを行うことで，ほとんどのデータについて一定の安全性が確保できると考えられる。今後，リスクレベルに関し社会的な合意を得た上で，安全な二次利用が促進されることが望まれる。

キーワード　院内がん登録，ミクロデータ，開示リスク，匿名化

Ⅰ　は　じ　め　に

わが国には，がんに関わる診療や国・地域の政策に役立てるためのがん情報を登録する仕組みとして，主に，「地域がん登録」「院内がん登録」「臓器がん登録」の３種類のがん登録がある。地域がん登録は，自治体（2011年12月１日現在45道府県１市）が実施主体となって運営しており，地域人口集団における全がんの罹患数・率，生存率を正確に把握することを目標とする。そのために個人情報を利用した名寄せを行い，予後情報を保持していることが特徴である。反面，進行病期に関する情報が臨床現場で用いられているものと異なり，現状では実施のない県も存在する。臓器がん登録は，学会や研究会による調査であり，各がん種に対応した詳細な臨床情報が利点ではあるが，症例が専門施設に偏っている可能性がある[1)2)]。

院内がん登録では，がん診療の連携拠点として全国で指定されているがん診療連携拠点病院（以下，拠点病院）において，各病院を受診したすべてのがん患者の部位・病理組織などの基礎データが集積されている。拠点病院の指定要件の１つとして，独立行政法人国立がん研究セ

＊１東京大学大学院医学系研究科公共健康医学専攻健康医療政策学分野博士課程　＊２同准教授
＊３独立行政法人国立病院機構北海道がんセンター臨床研究部長　＊４福井県立病院臨床病理科医長
＊５大阪府立成人病センターがん予防情報センター長　＊６栃木県立がんセンター研究所長
＊７群馬県立がんセンター院長　＊８独立行政法人医薬品医療機器総合機構主任専門員（調査・臨床医学担当）
＊９独立行政法人国立がん研究センターがん対策情報センターがん統計研究部診療実態調査室長
＊10同がん統計研究部長

ンターの主催する研修会を修了し，がんの病理組織，病期分類，院内がん登録のルールに関する一定の知識をもつがん登録実務者が当該施設に一人以上配置されることが定められており，彼らによってデータ収集が行われる。さらに，これらのデータは年に１度，国立がん研究センターがん対策情報センターに提供される。これは，2007年４月施行の「がん対策基本法」に基づく，「がん対策推進基本計画」による事業で，「がん診療連携拠点病院の整備に関する指針」（2008年厚生労働省健康局長通知）に従って行われている。個人情報の保護のため，各病院からデータが提供される際には，名前や住所の削除，生年月日を生年月に直すなどの処理が加えられ，直接的な個人情報を含まない状態となっている。提供されたデータは，2007年症例より１年ごとに集計と発表がなされている。2007年は287施設から316,089症例，2008年は353施設から420,683症例[3]が提供され，2008年からは施設別，施設名付きで集計値の報告がなされるようになった。これらは，各施設ががん診療の見直しや，地域の中での施設の役割を検討するために役立てることができる，がん診療の現状を示す貴重なデータであるといえる。

国立がん研究センターに提供されたデータは，現在は集計結果のみが発表されているが，この貴重なデータを最大限活用するために，将来的に，より多くの研究者などが解析利用できるようになることが望まれる。それには個人情報が守られる状態で，ミクロデータが利用できる体制を確立することが必要である。ミクロデータの提供・利用は，研究を促進するという意味では有用であるが，一方で，そこに含まれる情報から個人が識別され，その個人にとって知られたくない情報までが開示されるリスクを伴う。提供されたデータには，名前や住所など個人を直接的に示す項目は含まれないが，性別や年齢のように，第三者がある程度容易に知ることができる変数を組み合わせることで，個人が特定できる可能性があるからである。

本研究では，この院内がん登録全国データの開示リスク評価と，匿名化手法の検討を行った。開示リスクを減らすためには，制度的な利用者の制限と，データ情報量を減らすことで個々のデータと実際の個人が結びつかないようにするという２つの方法がある[4]が，本研究では後者，つまり匿名化の方法を検討する。そのために，収集されているデータを様々に加工した場合のリスクの変化を検討し，データ提供時の安全性を評価することを目的とする。

Ⅱ 方　　法

一般にリスクの評価には，標本一意，母集団一意という概念が用いられる[5]。標本一意は，性別や年齢のように第三者がある程度容易に知ることができ，間接的に個人の識別を可能とする変数（キー変数）の組み合わせによって，標本集団に含まれる個人が標本集団中で一人に定まる状態を指す。母集団一意は，標本集団に含まれる個人が母集団中で一人に定まる状態を指す。母集団で一意の場合，名前などの実際の個人情報がなくても理論上はキー変数の情報だけで個人が特定される。したがって，標本一意であり，かつ母集団一意である可能性が高ければ，そのデータはリスクが高いと考えられる。本研究ではこの考え方に基づくリスク評価を行う。

ここでの標本集団とは提供されるデータを構成する集団であり，母集団とは標本が由来する元の集団全体である。本研究の対象母集団は拠点病院を受診した全患者と考えると，院内がん登録データは拠点病院を受診した患者の全数調査であるから，基本的には標本集団＝母集団である。

院内がん登録データ2008年症例（424,983）を対象とし，年齢，性別，居住地，がんの部位に関する**表１**の変数をキー変数と考え，これらの組み合わせによる，一意症例数，一意症例割合を求めて，リスクの評価を行った。年齢，性別についてはデータの欠損はなかった。居住地については，診断時居住都道府県を用いることが妥当だが，欠損および不明のものが相当数存在する（欠損7,544，不明55）ため，欠損，不明のない診断病院所在都道府県で代用した。が

表1　本研究で検討したキー変数

個人識別につながる情報	データ内のキー変数
年齢	生年月
性別	性別
居住地	診断病院所在都道府県 診断病院名
がんの部位	部位（ICD-O-3の局在コードの数字上2桁。「胃」「気管支及び肺」など）

表2　匿名化の方法[6]

方法	説明
削除	直接的に個人識別につながる変数を削除
グルーピング	値を階級区分に変更。年齢を10歳刻みで表示等
トップ／ボトム・コーディング	個人識別につながりやすい特殊な属性をまとめる。年齢100歳以上等

表3　元データにおける識別リスク

	総数	%
計	424 983	100
一意	352 216	83
二意以上	72 767	17

んの部位については，第三者が容易に知ることができる情報は，詳細な亜部位ではなく，臓器分類程度と考え，ICD-O-3局在コードの数字上2桁を用い，部位不明は不明というグループとして他の部位と同様に扱った。

上記キー変数について，大域的再符号化と呼ばれる加工を行った。大域的再符号化とは，データ全体に対して一律に情報量を減らす，例えば全データについて1歳刻みの年齢データを5歳刻みにグルーピングする，などの処理を意味する。性別，居住都道府県，部位の情報量を減らすことは，データの有用性を大きく損なうため，診断病院名の削除，年齢の5歳刻みおよび10歳刻みのグルーピング，トップ／ボトム・コーディング[6]を行った（**表2**）。年齢のトップ／ボトム・コーディングの閾値は，総務省統計局の研究会の報告書で「海外では，トップ・コーディングされるのが対象全体の0.5%以上としている例などがある」と記載されている[6]ことに倣って，その年齢未満，以上の症例がそれぞれ0.5%以上となる20歳未満，90歳以上とした。

各変数の大域的再符号化をそれぞれ単独で，また組み合わせて行った後，再び一意数，一意割合を求めてリスクの再評価を行った。ある程度の一意の割合が低下して，安全性が担保されると考えられるまで，匿名化と評価を繰り返した。また，居住地の情報量を，都道府県から7地域（北海道・東北，関東，中部，近畿，中国，四国，九州・沖縄）に減らした場合についても検討した。

大域的再符号化によってある程度の安全性が担保された後，さらに安全性を上げる手法としてリサンプリングの効果も検討した。リサンプリングとは，データの中から一定の割合を抽出するものである。ユーザーからすると，より小さい標本集団で調査が行われたのと同じである。標本集団が小さくなると，標本一意であっても，それが母集団一意である確率が下がるため，安全性が向上する。

なお，本研究は，国立がん研究センターおよび東京大学にて倫理審査を受け，承認を得ている。

Ⅲ 結　　果

元データにおける，キー変数によって定まる一意数，一意割合を**表3**に示す。特に匿名化の処理を施さない元データでは，全体の83%が標本一意かつ母集団一意であった。

大域的再符号化の結果は**図1**となる。一意割合は，診断病院名の削除のみでは46%，年齢のグルーピングとトップ／ボトム・コーディングのみでは9～14%であったが，それらを組み合わせることで2～3%にまで低下した。一方で，7,140～12,699例が依然として一意であった。さらに都道府県を地域にまとめると一意割合は0.2%，一意数は1,006にまで低下した。

診断病院名の削除および年齢の10歳刻みでのグルーピングとトップ／ボトム・コーディングを行ったデータ（**図1**の診断病院名削除＋年齢

図１　大域的再符号化による識別リスクの減少

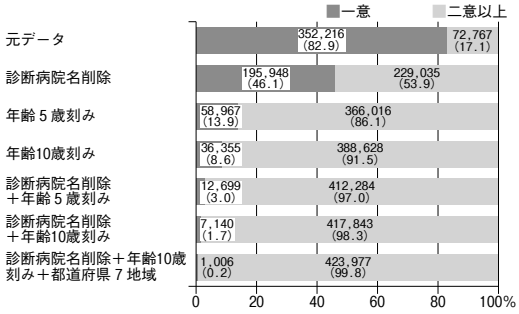

表４　診断病院名削除＋年齢10歳刻みの処理後からのランダムサンプリングにおける抽出率と一意数

抽出率（%）	抽出数	標本一意数①	うち母集団一意数②	②／①（%）
0	0	0	0	0.0
10	42 498	5 000	714	14.3
20	84 997	5 938	1 429	24.1
30	127 495	6 417	2 141	33.4
40	169 993	6 703	2 854	42.6
50	212 492	6 888	3 570	51.8
60	254 990	7 003	4 282	61.2
70	297 488	7 074	4 999	70.7
80	339 986	7 115	5 712	80.3
90	382 485	7 132	6 424	90.1
100	424 983	7 140	7 140	100.0

10歳刻みの状態）において，グループ化によるさらなる秘匿が可能であるかを検討するために，これら匿名化処理後も一意であるデータの特徴を記述したが，すべてのキー変数について均一な分布を持っており，際だった特徴がなかったため，さらなるグループ化による匿名化は効果的ではないと判断した。

そこで，次の段階として同じく**図１**の診断病院名削除＋年齢10歳刻みの状態からリサンプリングを行った場合の結果は，**表４**のようになる。抽出率を変えたランダムサンプリングをそれぞれ1,000回行った結果である。例えば50％の抽出率では，抽出後の標本で一意に見えるデータの約半数が母集団一意でなかった。

Ⅳ 考　　察

本研究においては，拠点病院から提供されている院内がん登録の全国データについて，データ内容に共通のない一意症例の割合から匿名化手法の開示リスク検討を行った。診断病院名の削除に加えて，年齢の５歳または10歳刻みでのグルーピングとトップ／ボトム・コーディングを行うことで，ほとんどのデータについてある程度の安全性が確保できると思われる。また，処理後も一意であるデータがキー変数について均一な分布を持っているため，大域的再符号化で，これ以上に識別リスクを減少させるためには，全体的により粗いグルーピング（年齢の刻みを荒くする，居住地を都道府県ではなく地域にするなど）をする必要があり，データの有用性を損なう可能性が高い。よって，別の手法を併用する必要がある。

それ以上の安全性のためには，リスクの高いデータのみへの局所再符号化としての局所欠損化やスワッピング，あるいはリサンプリングなどを行う必要がある。局所欠損化とは，そのデータの特定の変数を欠損させること，スワッピングとは，類似のデータ間で，一部の変数の値を入れ替える方法である。リサンプリングは，現存するデータの中から一定の割合を抽出するものであり，本研究ではこの影響も検討した。結果で示したとおり，抽出後の標本で一意であっても母集団一意である割合が低下し，どのレコードが真に一意であるかを不明とする有用な方法である。米国の政府統計の公開でもリサンプリングは広く用いられており，原調査の抽出率によって異なるものの，規模の大きい行政登録データでは１～数％，出生登録レジスターでは50％の抽出率で提供されている[7)]。

どの程度まで匿名化を施せば安全とするかは，制度的に議論すべき内容である。例えば最大限大域的符号化を施した後の0.2％の一意割合は微小であると思える反面，実数では1,000名以上の一意数があり，これらの人数が開示リスクにさらされているとも考えることができる。しかし，本研究で示したように，データの有用性を損なわない程度の大域的再符号化とリサンプリングを行えば，作成されたデータにおいて一意であっても，そのうちのどれが母集団一意かわからない状態になることから，安全性が確保

されていると考えることができる。そのため，将来的にはこれらの処理を施したデータは，二次利用目的で広く提供できる可能性が高いと思われる。また，院内がん登録のデータは毎年国立がん研究センターでデータの集積が継続的になされていくため，今後具体的なデータ提供に当たって同様の手法を用いて開示リスクの客観的評価を行い，提供判断を行っていくことが可能と考えられる。

政府統計におけるミクロデータの開示は欧米諸国では幅広く行われており，例えばアメリカでは1960年代から様々なミクロデータが公開されて，その利用により多くの成果が得られている[7]。日本は，旧統計法でその点で立ち遅れてきたが，2009年全面施行の新統計法で二次的利用制度が創設され，一般学術研究に個票データ利用の道が開けただけでなく，オーダーメイド集計や匿名データの作成，提供など，公共の財産である政府統計データを有効利用する方向に進みつつある。院内がん登録データも，公共性の観点からは政府統計に準じる貴重なデータであるため，今後も，将来的なデータの公開を見据えた検討を進めていく必要がある。

院内がん登録全国データは，わが国のがん対策や臨床医学の発展のため広く利用されるべきである。その際に，情報の有用性と安全性のバランスは必須であろう。解析の目的は多様なものが考えられ，必要になるデータ粒度も解析目的に応じて変化させる必要がある。本研究は，一定の条件における匿名化の安全性の評価であるが，今後の多様な要望に応じた柔軟な安全性評価にも応用できる可能性があると考えられる。

謝辞

本研究を行うにあたり，ご協力下さいましたすべての方々に深謝致します。温かいご指導をいただきました。東京大学情報理工研究科の竹村彰通教授，岡山商科大学経済学部の佐井至道教授，金沢大学経済学部の星野伸明准教授，東京大学大学院医学系研究科公共健康医学専攻の小林廉毅教授，豊川智之講師，冨尾淳助教と大学院生の皆様に深く感謝申し上げます。

なお，本研究は第３次対がん総合戦略事業「院内がん登録の標準化と普及に関する研究」の助成を得て行われた。

文　　献

1）東尚弘，祖父江友孝，西本寛．臓器がん登録の現状－臓器がん登録の実態についての調査報告－．外科治療 2011；104（２）：169-76.

2）猿木信裕，編．がん登録の軌跡．東京：悠飛社，2010.

3）独立行政法人国立がん研究センターがん対策情報センターがん統計研究部院内がん登録室，編．がん診療連携拠点病院　院内がん登録　2008年全国集計報告書．2011.

4）Willenborg, L. and T.d. Waal, eds. Lecture Notes in Statistics, Elements of Statistics of Disclosure Control. Berlin / Heidelberg: Springer, 2001.

5）竹村彰通，個票開示問題の研究の現状と課題．統計数理 2003；51（２）：241-60.

6）総務省統計局統計データの二次利用促進に関する研究会，統計データの二次利用促進に関する研究会報告書．2008.

7）松田芳郎，濱砂敬郎，森博美．講座　ミクロ統計分析　統計調査制度とミクロ統計の開示．東京：日本評論社，2000.