研究成果申告書（がん対策推進総合研究事業）

<table>
<tr><td colspan="4">研究課題名　全国がん登録の円滑な運用のための検証に関する研究　（20EA1007）</td></tr>
<tr><td colspan="4">研究代表者名　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）</td></tr>
<tr><td colspan="2">研究分担者の人数：　４人</td><td colspan="2">研究期間：令和２年度　～　令和４年度</td></tr>
<tr><td rowspan="2">研究費交付額（追加交付を含む）
（単位：千円）</td><td>令和２年度</td><td>令和３年度</td><td>令和４年度</td></tr>
<tr><td>12,000 千円</td><td>9,790 千円</td><td>9,790 千円</td></tr>
<tr><td rowspan="2">評価点数（単位：点）</td><td>事前評価</td><td>中間評価
１年目</td><td>中間評価
２年目</td></tr>
<tr><td>14.1 点／20 点</td><td>15.3 点／20 点</td><td>X.X 点／20 点</td></tr>
</table>

がん登録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。

前者①は、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。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の手がかりが得られ、この経時的評価で安定化の過程がわかる。情報内容の質については、部位・組織型の詳細度（NOS の頻度など）や部位と組織型の分布など情報自体の特性を検討することで評価が可能である。これらは 1～3 年次各年で行いその動きを検証する。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。

後者②データ匿名化の安全性評価の確立については、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。研究者への提供では、研究課題によっては全データの提供を希望する課題もあり、データの提供ではなく遠隔解析システムの導入などが安全性確保のために必要となる。公表における少数例秘匿については常に有用性とのトレードオフ関係にあり、その適正な評価方法も開発しなければならない。こちらは、初年は海外やわが国の他統計基準の状況をまとめて方針計画を立て、2 年目で制度応用に必要な解析を行って 3 年目で提言をまとめる。

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

求められる成果 ・匿名化された情報の提供の手法の検討 ・全国がん登録システムの運用方法の検証 ・匿名化された全国がん登録の匿名化指標の開発

目標・成果物	（期待される効果） ①　匿名化された情報の提供の手法の検討 ①-1　匿名化を破る攻撃者が持ちうる情報など考え、想定を作成する ①-2　全国がん登録 2016 年、2017 年の匿名化データを申請する ①-3　全国がん登録データを用いて 1 で考えられる項目を加工して k-匿名化で評価する ②　匿名化された全国がん登録の匿名化指標の開発 全国がん登録の匿名化の強度を測るための指標を同定する。あるいは匿名安全性を守るための客観的基準を同定する。 ②-1　k-匿名化をもとに応用した匿名化指標あるいは安全性基準を検討する ②-2　個票データの安全性だけではなく、集計データの安全性の評価と基準を検討する ②-3　外国におけるデータ提供先行事例のモデルを選定し、対象を絞って詳細な情報収集を行う ③　全国がん登録システムの運用方法の検証 予後の妥当性を検証するために 2016 年症例の 3 年予後を全国がん登録からのデータと従来からの住民票照会によるデータを比較検討する ③-1 国立がん研究センター中央病院の 2016 年症例についてこれまでの予後調査支援事業に倣い住民票照会を行う ③-2 2016 年症例の 3 年予後の確定時点でがん登録推進法 20 条に基づく生存確認情報の提供を受ける ③-3 住民票照会結果と全国がん登録提供を受けた生存確認情報を比較する ③-4 国立がん研究センター中央病院以外の症例収集の可能性の検討 ④　全国がん登録の運営上で算出される指標を設定し、制度移行の影響などについてのモニタリングを行う
目標・成果物の達成状況（1 年目）	**①　匿名化された情報の提供の手法の検討［達成見込み（令和 5 年 3 月）］** 全国がん登録の実際のデータを用いて、様々な加工を行うことで k-匿名化（同一のデータが k 人以内となる状況）の達成度について k を変化させながら検討する。

	①-1　匿名化を破る攻撃者が持ちうる情報など考え、想定を作成する ［達成済み］ 攻撃者の立場を、知人、医療者など、さまざまな想定をしつつ、それらの立場の人間が知りうる情報を全国がん登録の提供項目から同定し、攻撃パターンを検討した。 ①-2　全国がん登録２０１６年、２０１７年の匿名化データの利用の申出を行う［達成済み］ 11 月の匿名化情報の提供に関して申出を行い、審議会でデータ利用が認められた。 ①-3　全国がん登録データを用いて１で考えられる項目を加工して k-匿名化で評価する［未達成］ データが到着したばかりのため、これから解析を開始するところである。 **②　匿名化された全国がん登録の匿名化指標の開発［達成見込み（令和 5 年 3 月）］** 全国がん登録の匿名化の強度を測るための指標を同定する。あるいは匿名安全性を守るための客観的基準を同定する。 ②-1　k-匿名化をもとに応用した匿名化指標あるいは安全性基準を検討する ［未達成・達成見込み（令和 4 年 3 月）］ ②-2　個票データの安全性だけではなく、集計データの安全性の評価と基準を検討し、一定の提言をする ［未達成・達成見込み（令和 5 年 3 月）］ ②-3　外国におけるデータ提供先行事例のモデルを選定し、対象を絞って詳細な情報収集を行う 匿名化強度の評価方法について情報収集する過程において、文献でカナダ・オンタリオ州において具体的に漏洩リスクを、理論的にパラメーラを用いて推定する手法が紹介されていたことから、オンタリオ州の公的データ取り扱い機関である ICES(Institute of Clinical Evaluative Science)の情報収集を開始した。今後、機関としての体制や組織発展の経緯などの詳細を情報収集し、まとめを行う[達成見込み（令和 3 年 6 月頃）] **③　全国がん登録システムの運用方法の検証［達成見込み（令和 5 年 3 月）］** 予後の妥当性を検証するために 2016 年症例の 3 年予後を全国がん登録からのデータと従来からの住民票照会によるデータを比較検討する ③-1 国立がん研究センター中央病院の 2016 年症例についてこれまでの予後調査支援事業に倣い住民票照会を行う［達成見込み（令和 3 年 3 月）］

	2016 年症例のうち①死亡患者全件、②死亡が確認されない患者のうち、生存最終確認日が 2019/12/31 以前の者で、バイオバンク（予後追跡）非同意患者様除いた 3824 件について予後調査支援事業と同等の方法で、各地へ住民票照会を行った。12 月現在作業中である。 ③-2 国立がん研究センター中央病院 2016 年症例の 3 年予後の確定時点でがん登録推進法 20 条に基づく生存確認情報の提供を受ける［未達・達成見込み（令和 4 年 6 月）］ 3 年予後の確定は 2019 年の生存情報確定となるため、2021 年（令和 3 年中）に予定されている。それが使えるようになるのは令和 4 年度となる。 ③-3 住民票照会結果と全国がん登録提供を受けた生存確認情報を比較する［未達・達成見込み（令和 5 年 3 月）］ 上記③-2 の作業が終わってからになる。 ③-4 国立がん研究センター中央病院以外の症例収集の可能性の検討 ［未達・達成見込み（令和 4 年 3 月）］ 当初想定していた、がん登録推進法 20 条で施設へ提供された生存確認情報は第 3 者提供ができず、院内がん登録実施に係る指針に記載のある生存率の全国集計のみ国立がん研究センターが収集することが可能とされている。そこで、国立がん研究センター中央病院のデータのみを使った検証を開始したが、本研究の目的で全国がん登録の運営のために収集したり突合したりすることが可能かどうか、については未整理である。今後可能な方法についての検討を継続的に行う必要があると考えている。 **④　全国がん登録の運営上で算出される指標を設定し、制度移行の影響などについてのモニタリングを行う［未達成・達成見込み（令和 5 年 3 月）］** 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指標の設定と算出可能性について検討中である。こられの各種指標については、一部は、全国がん登録システムの運用監視の中で画面上管理が可能となっているが、より詳細な突合診断年分布などの集計については、システム上新たな抽出集計設定が必要なものが多数含まれることが判明したため引き続き検討中である。また、このような内容の外部への報告・公表については、審査の必要なデータの利活用にあたるのかなどの整理が必要であるため、現段階では本研究としては検討にとどまっている。今後、令和 3 年 3 月までに道筋としての整理を行い、必要であれば審議委員会への申出を行う。
目標・成果物の	（期待される効果） **①　匿名化された情報の提供の手法の検討**

達成状況（2 年目）	①-1　匿名化を破る攻撃者が持ちうる情報など考え、想定を作成する（１年目達成済み） ①-2　全国がん登録 2016 年、2017 年の匿名化データを申請する（１年目達成済み） **①-3　全国がん登録データを用いて１で考えられる項目を加工して k-匿名化で評価する**［未達・達成見込み（令和 4 年 3 月）］ データの秘匿については、全国がん登録 2016 年症例、17 年症例の利用申出を行い、データの提供を受けた。基本的な安全性の確認のため、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を使用した時、さらに性別、年齢を組み合わせた時の k-匿名化の度合いを知るために、ユニーク（K=1）となる症例を集計した。また、逆に k-匿名化を目標にデータを加工していくプログラムを作成している。もう少し解析を追加したのちに、公表確認の上で報告とする。また、地理情報の粒度による k-匿名化のレベルを検討するために、市区町村、二次医療圏、保健所コード毎、都道府県でユニークとなる地域の数を検討した。 **②　匿名化された全国がん登録の匿名化指標の開発** 全国がん登録の匿名化の強度を測るための指標を同定する。あるいは匿名安全性を守るための客観的基準を同定する。［未達成・達成見込み（令和 4 年 8 月ごろ）］ ②-1　k-匿名化をもとに応用した匿名化指標あるいは安全性基準を検討する［未達成・達成見込み（令和 4 年 8 月ごろ）］ k-匿名化のみではなく、l-多様性等の他の評価方法について検討中である。さらに、逆に k-匿名化を確保したデータを作成するための R 言語によるルーチンの開発に着手した。 ②-2　個票データの安全性だけではなく、集計データの安全性の評価と基準を検討する［未達成・達成見込み（令和 5 年 3 月ごろ）］ 集計データとして地理情報が公表された場合の安全性について検討するために、国勢調査のデータの二次利用を申請して、現在審査中である。 **③　全国がん登録システムの運用方法の検証** 予後の妥当性を検証するために 2016 年症例の 3 年予後を全国がん登録からのデータと従来からの住民票照会によるデータを比較検討する ③-1 国立がん研究センター中央病院の 2016 年症例についてこれまでの予後調査支援事業に倣い住民票照会を行う［未達成・達成見込み（令和 4 年 3 月ごろ）］ 2016 年症例については前年度で実施済み。調査が完了し 7094 名の対象者のうち通院中あるいは追跡に同意の得られていない 3270 名を除き、3824 名を照会、3749 名について住民票の取得が可能であった。結果、生存 1333 名、

	死亡 3849 名、148 名については不明という結果を得た。7351 名の対象者のうち 3347 名の通院者、非同意者を除き 4004 名を対象として住民票照会を実施した。 ③-2 3 年予後の確定時点でがん登録推進法 20 条に基づく生存確認情報の提供を受ける［未達・達成見込み（令和 4 年 8 月）］ 前項③-1 が完了してから、令和 4 年度の半ばごろにとなる。 ③-3 住民票照会結果と全国がん登録提供を受けた生存確認情報を比較する［未達・達成見込み（令和 5 年 3 月）］ 前項③-2 が完了してから、解析が可能になる。 ③-4 国立がん研究センター中央病院以外の症例収集の可能性の検討［未達・達成見込み（令和 5 年 3 月）］ 現行法制化においては、がん登録推進法 20 条により各施設に提供されたデータを院内がん登録の実施にかかる指針に記載された目的以外に収集するのは困難であるが、引き続き検討する。 **④ 全国がん登録の運営上で算出される指標を設定し、制度移行の影響などについてのモニタリングを行う**［未達・達成見込み（令和 5 年 3 月）］ 診断施設不明例の経過を全国がん登録の集計の中で追跡し、2016 年症例は 69141（7.0%）、2017 年症例は 59606（6.1%）2018 年症例は 54489（5.6%）と漸減傾向を観察した。また、全体的な他の指標の検討については引き続き検討する。
目標・成果物の達成状況（3 年目）	（期待される効果） **① 匿名化された情報の提供の手法の検討** ①-1 匿名化を破る攻撃者が持ちうる情報など考え、想定を作成する（1 年目達成済み） ①-2 全国がん登録 2016 年、2017 年の匿名化データを申請する（1 年目達成済み） **①-3 全国がん登録データを用いて 1 で考えられる項目を加工して k-匿名化で評価する**［達成済み（令和 4 年 10 月）］ データの秘匿については、全国がん登録 2016 年症例、17 年症例の利用申出を行い、データの提供を受けた。基本的な安全性の確認のため、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を使用した時、さらに性別、年齢を組み合わせた時の k-匿名化の度合いを知るために、ユニーク（K=1）となる症例を集計した。その結果、2 か年を合わせたところでは、ICD-O-3 の部位分類のみでのユニークになるものは 58 件しかなく、ICD-10 分類でも 86 件だったのに対して、部位組織分類まで含めるとは 4639

	件のユニーク症例が出現した（資料 1，表 1)。これらを削除したとしても全体の件数は 200 万件以上ある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る。一方で、ICD-O-3 の部位・組織分類と性別、年齢（1 歳刻み）を加えると 87688 件のユニークなレコードが生まれた。年齢を 5 歳刻みにすると 34800 までおさえられ、また、ICD-10 分類では、2087 件まで抑えられた（資料 1，表 1)。これらのトレードオフについては研究仮説とその有用性に応じて検討をする必要があると考えた。 **② 匿名化された全国がん登録の匿名化指標の開発** 全国がん登録の匿名化の強度を測るための指標を同定する。あるいは匿名安全性を守るための客観的基準を同定する。 ②-1 k-匿名化をもとに応用した匿名化指標あるいは安全性基準を検討する k-匿名化［達成済（令和 4 年 9 月）］ k-匿名化を確保したデータを作成するための R 言語によるアプリを検討した。ここでは、通常の都道府県、市区町村、町丁目といった地域レベルに基づく地域情報の匿名化処理に加え、地域の位置座標（緯度、経度）に基づき柔軟に地域領域を再帰的に分割する k-匿名化アルゴリズムを開発し実装した。全国がん登録に含まれるレコードを開発した手法で「診断時患者住所」を k-匿名化し、従来の地域レベルに基づく k-匿名化処理に比べ、分割したグループの粒度の均一化に関する有用性指標が 5%から 15%改善したことを実証的に評価した（資料 2)。 さらに既存の匿名化ツール ARX を用いて、患者の性別、年齢、都道府県コードを準識別子、ICT-10 を機密情報と想定し、k-匿名化したデータに対してより厳しい属性推定のリスクを考慮した安全性指標である l-多様性で評価した。その結果、k-匿名化に基づくレコードのグループ化では ICT-10 の異なる値の数が k 未満となり十分な多様性が維持できない場合があることが判明した。また l-多様性, t-近似性を満たす匿名化データの有用性評価を行ったが、それらの安全性を満足するには通常の k-匿名化よりもより粒度の粗いデータに加工する必要があり、特に t-近似性の場合は大幅に匿名化データのグループサイズが大きくなることが判明した。今後もプライバシー保護とデータ有用性のトレードオフについては検討する必要があることが明らかになった（資料 2)。 ②-1’ これらの知見をもとにまた、文献や海外事例を総合して、匿名化指標あるいは安全課基準を提案する。[未達成・達成見込み（令和 5 年 3 月ごろ）] ②-2 個票データの安全性だけではなく、集計データの安全性の評価と基準を検討する［一部未達。達成見込み（令和 5 年 3 月ごろ）］ 集計データとして地理情報が公表された場合の安全性について検討するために、国勢調査のデータの二次利用を申請し、全国（母集団）の市町村の年齢/性別構成を確認した。同時に、全国がん登録情報より、がん患者の同条件

	の構成を母集団と比較検討した。 まず 5 歳階級別に、市町村別がん患者数、及び、がん患者が 1 人になる市町村数を算出した。その結果、どの年齢階級においても、がん患者が 1 人となる市町村が存在した（資料 3，表 1-1)。特に、20 歳未満と 100 歳以上の場合、当該年齢階級のがん患者が居住する市町村のうち半数以上の市町村において、がん患者が 1 人となっていた（資料 3，表 1-2)。さらに男女別でみると、これらの傾向はさらに強まっていた（資料 3，表 2-1～3-2)。また、年齢構成を 10 歳階級に条件を変えて同様の集計を行ったが、当該年齢階級において、がん患者が 1 人となる市町村の割合は 2 割程度減るものの、同様の傾向を示した（資料 3，表 4-1～6-2)。 以上より、患者の地理情報を市町村まで提供する場合、年齢を 5 歳階級別に加工して提供したとしても、市町村に当該年齢階級のがん患者が 1 人となるケースが多く存在するため、個人識別リスクが高い状態であることが示唆された。また、通常、性別も同時に提供することが多いため、性別が特定可能な場合は、さらに個人識別リスクが高まることが推察される。たとえ年齢を 10 歳階級に加工して提供としたとしても、個人識別性が低くなるとは言い切れず、市町村レベルの患者住所の提供を行う際は、個人識別リスクについて十分検討が必要である。 今後、さらなる解析を行い、がん登録情報の提供関連の安全性評価を行う。 **③　全国がん登録システムの運用方法の検証** 特に予後情報が今後重要になると考えられるが、予後の妥当性を検証するために 2016 年症例の 3 年予後を全国がん登録からのデータと従来からの住民票照会によるデータを比較検討する ③-1 国立がん研究センター中央病院の 2016、2017 年症例についてこれまでの予後調査支援事業に倣い住民票照会を行う［達成済み（令和 4 年 3 月）］ 2016 年症例については住民票照会を実施済み 2017 年症例についても同様の紹介作業を実施した。 ③-2 3 年予後の確定時点でがん登録推進法 20 条に基づく生存確認情報の提供を受ける［未達・達成見込み（令和 4 年 12 月）］ 国立がん研究センター中央病院において生存確認情報の提供を東京都から受けたことを確認し、次に、施設内における院内がん登録利用申請を現在準備中である。 ③-3 住民票照会結果と全国がん登録提供を受けた生存確認情報を比較する［未達・達成見込み（令和 5 年 3 月）］ 前項③-2 が完了してから、解析が可能になる。 ③-4 国立がん研究センター中央病院以外の症例収集の可能性の検討［未達・

	達成見込み（令和５年３月）］ 現行法制化においては、がん登録推進法 20 条により各施設に提供されたデータを院内がん登録の実施にかかる指針に記載された目的以外に収集するのは困難であるが、引き続き検討する。 **④　全国がん登録の運営上で算出される指標を設定し、制度移行の影響などについてのモニタリングを行う**［未達・達成見込み（令和５年３月）］ 診断施設不明例の経過を全国がん登録の集計の中で追跡し、2016 年症例は 69141（7.0%）、2017 年症例は 59606（6.1%）2018 年症例は 54489（5.6%）、さらに、2019 年症例は 49,482（5.0%）と漸減傾向を観察した。最新年であっても 5～6%の減少傾向を認めており、まだ精度は今後向上していくものと考えられる。全体的な他の指標については、届出総数などが考えられるが、令和４年３月のシステムの更新によって 2021 年症例から検討が可能になると予想される。

目標・成果物の達成状況を証明する資料集

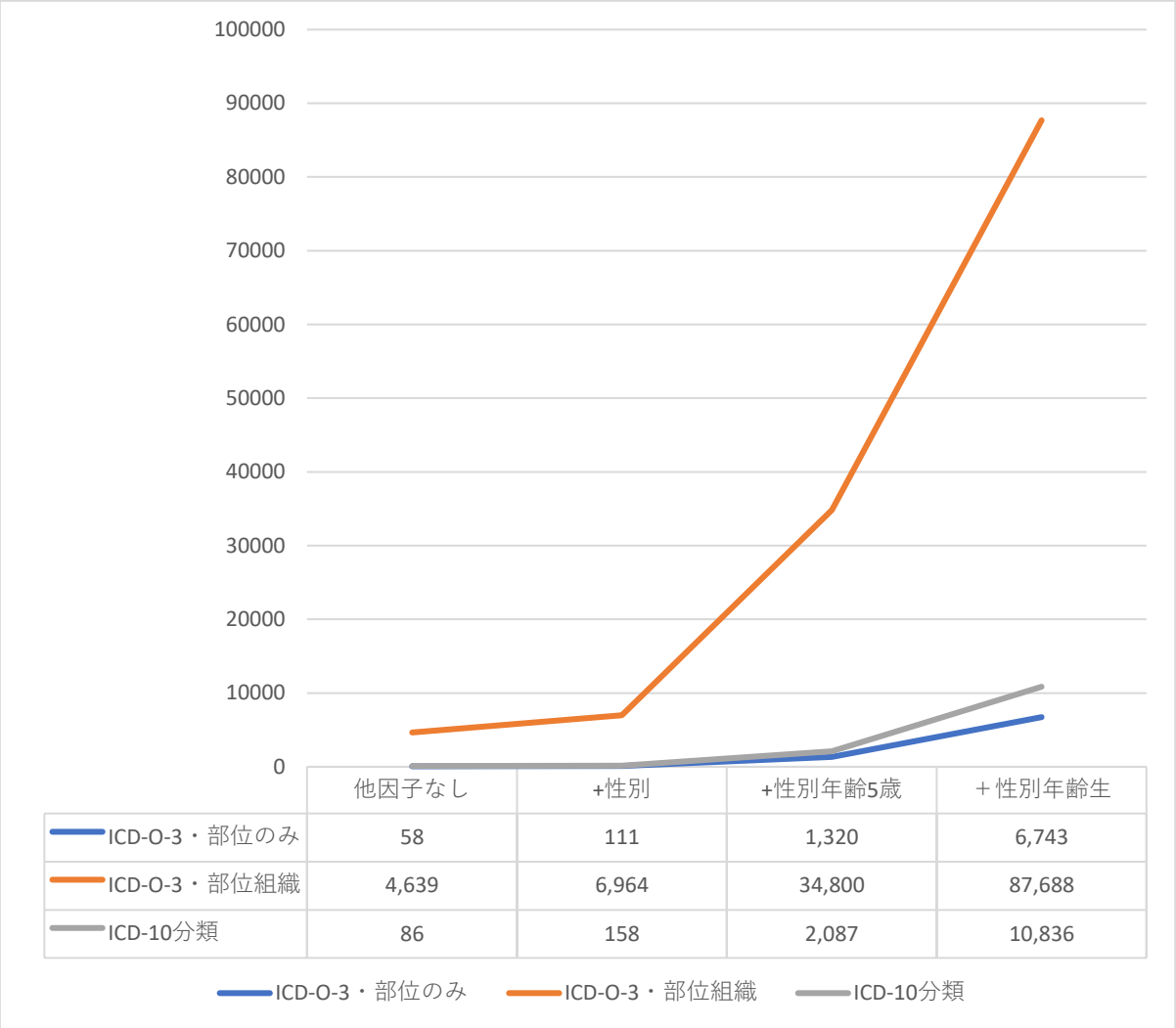
資料 1　匿名化攻撃者による観察可能項目の評価

資料 2　GPS 位置情報に基づく地域情報の再帰的な匿名化処理とその実証評価
（2022 年度統計関連学会連合大会　発表資料）

資料 3　地理情報が公表された場合の安全性についての評価

資料１　匿名化攻撃者による観察可能項目の評価

表 1. 2016 年及び 2017 年の 2 か年で、ICD コード（ICD-O-3）、それに順次、性別、5 歳刻み年齢、1 歳刻み年齢を組み合わせた場合のユニークになる票（患者）数



	他因子なし	+性別	+性別年齢5歳	＋性別年齢生
ICD-O-3・部位のみ	58	111	1,320	6,743
ICD-O-3・部位組織	4,639	6,964	34,800	87,688
ICD-10分類	86	158	2,087	10,836

資料 2　GPS 位置情報に基づく地域情報の再帰的な匿名化処理とその実証評価

（2022 年度統計関連学会連合大会　発表資料）

GPS位置情報に基づく地域情報の再帰的な匿名化処理とその実証評価

統計数理研究所
南 和宏

2022年9月8日

2022年度統計関連学会連合大会
企画セッション「大規模データにおける匿名加工とプライバシー保護の新たな展開」

1

本日の内容

- がん登録情報の匿名化の課題
- 国土交通省 位置参照情報とのリンケージ
- Mondrianアルゴリズムによる地域情報のk-匿名化処理
- 匿名データの有用性評価
- 今後の課題

今回利用したがん登録情報、法に基づき情報の提供を受け、独自に作成・加工した資料等である。

2

がん登録情報

https://ganjoho.jp/public/institution/registry/national.html

- 全国のがん患者の情報を全国がん登録データベースとして、国（国立がん研究センター）が一元管理
- 患者単位のミクロデータであり、2,324,132レコード、69項目からなる。
  - がんと診断された人の氏名、性別、生年月日、住所
  - がんの診断を行った医療機関名
  - がんの診断を受けた日
  - がんの種類

3

公的統計における匿名データ

- 一般の利用に供することを目的として調査票情報を特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む。）ができないように加工したものをいう。(統計法第2条第12項)

レコード識別の防止が要件

4

匿名化の基準

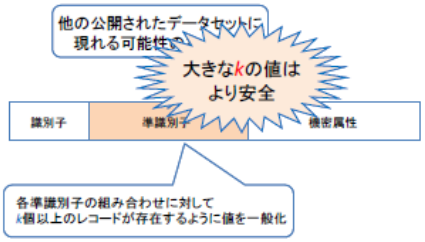
- 調査票情報は、統計調査の目的や規模等によってその特性が異なり、一律に匿名化の基準を設定することは困難であることから、これまで提供機関は、匿名化する統計調査の特性を勘案し、匿名化の基準となる値(例えば、年齢、世帯人員などの識別情報のトップコーディングに当たって母集団全体の 0.5%とすること等)を個別に定めてきた。

匿名データの作成・提供に関するガイドライン

5

k-匿名化はレコード識別防止を一般的した概念

アイデア：必ず同一の準識別子の値を取るレコードをk個以上作成し、レコード識別をk個未満に絞り込ませない

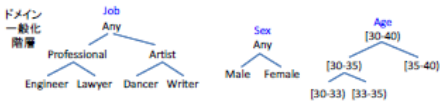


6

例: 3-匿名化テーブル

準識別子(一般化の対象)　機密属性(値は変えない)

	Job	Sex	Age	Disease
等価クラス	Professional	Male	[35 – 40)	Hepatitis
	Professional	Male	[35 – 40)	Hepatitis
	Professional	Male	[35 – 40)	HIV
	Artist	Female	[30-35)	Flu
	Artist	Female	[30-35)	HIV
	Artist	Female	[30-35)	HIV
	Artist	Female	[30-35)	HIV



7

k-匿名化の実施上の課題

- 安全性パラメータkをどう選ぶか？
  - kの値が大きければ安全になるが、情報損失は増大
- どの変数(属性)を準識別子とするか？
  - 攻撃者の外部知識を適切に想定する必要がある
- k-匿名化データ作成のためにどのような一般化処理(グループ化)を行うか？

がん登録情報に適した匿名化手法を検討

8

攻撃者と準識別子

- 攻撃者は患者の近所の知人を想定
- 準識別子として、今回は患者の住所情報を特に検討
  - 診断時患者住所
    - 県レベル
    - 市レベル
    - 町レベル
    - 丁目レベル
  - 性別
  - 診断時年齢

今回は住所情報のみを用いてk-匿名化の等価クラスを作成・評価

9

地域情報の一般化処理の問題点

- 地域間の人口密度の違いは大きく、地域レベルによる一般化では、等価クラスの粒度を均一にするのが困難

GPS座標に基づく、
地域領域の柔軟な分割

10

国土交通省 位置参照情報

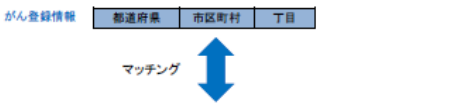
(https://nlftp.mlit.go.jp/cgi-bin/isj/dls/_choose_method.cgi)

- 全国の都市計画区域相当範囲を対象に、大字・町丁目レベル(大字・町丁目の住所代表点)の位置座標を提供
  - 都道府県コード
  - 都道府県名
  - 市区町村コード
  - 市区町村名
  - 大字町丁目コード(JIS市区町村コード＋独自7桁)
  - 大字町丁目名
  - 緯度(単位：度、小数点以下第6位まで、半角)
  - 経度(単位：度、小数点以下第6位まで、半角)
  - 原典資料コード(1：自治体資料、2：街区レベル位置参照情報、3：1/25000地形図、0：その他資料)

11

位置参照情報とのリンケージ

- 都道府県、市区町村、丁目の値が同じレコードをMySQLで連結



12

リンケージの課題

- 大字・町丁目レベルの地名が2つのデータセットで異なる場合がある
  - 位置参照情報に含まれる地名は、市町村資料、国土地理院の25000分の1地形図、民間の地図等を基に作成しており、国内の標準的な地名を指定しているものではない
- がん登録情報における市区町村、丁目に欠損値の存在

13

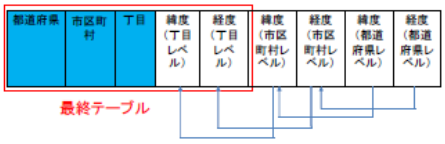
解決策

- 位置参照情報から都道府県レベル、市区町村レベル、丁目レベルのGPS座標の3種類のテーブルを作成
- MySQL の RIGHT JOIN でがん登録情報の住所情報（以下、住所テーブル）をRight table に設定
  - 住所テーブルのレコードはすべて保持
- 住所テーブルを3種類のGPSテーブルとRIGHT JOINを実行

14

解決策（続き）

- 市区町村レベルの欠損は、都道府県レベルの座標値を補完
- 丁目レベルの欠損は、市区町村レベルの座標値を補完



15

今回はデータの効率的な処理のため、サマライズしたテーブルを使用

- 同一の（都道府県，市区町村，丁目）をもつレコードの数のカラムを追加し、同一レコードの繰り返しを省略

Original microdata

pref	city	machi	chome
Tokyo	A	B	C
Tokyo	A	B	C
Tokyo	A	B	C

変換

Summary table

pref	city	machi	chome	Size
Tokyo	A	B	C	3

＜表データの形式＞

都道府県	市区町村	丁目	緯度（丁目レベル）	経度（丁目レベル）	レコード数

16

k-匿名化データの作成

- R言語でGPS座標値に対するMondrianアルゴリズムをR言語で実装
  - 地域情報のトップダウンの再帰的分割により、柔軟な地域グループの作成を目指す
- 安全性パラメータを *k=500,1000,2000* に設定し、住所情報の等価クラスを作成
  - 小地域による匿名化の可能性を試行
- データの有用性をDiscernibility指標で評価
  - 地域レベルの一般化による匿名化手法との比較

17

Mondrian [LeFevre06]

- トップダウン型のどん欲(Greedy)アルゴリズム
- *k*-匿名化を領域分割問題として定式化
- 事前にドメイン階層を定義する必要がない
- Discenability指標 (C_{DM})と呼ぶペナルティの最小化を目指す

$$C_{DM} = \sum_{G_i} |G_i|^2$$

- つまり、出来るだけ均等なサイズ*k*のグループに分割しようとする

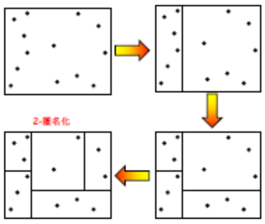
18

Mondrianアルゴリズム

- 今回の基準点は、領域内のGPS座標の中央値を使用
- 東西、南北の方向に交互に分割

入力：領域L
　　領域上の点集合S
　　しきい値k
出力：部分領域の集合

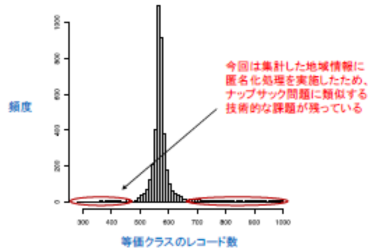
1.領域Lを分割する次元（属性）を決める
2. その次元を軸に中間値で分割
3. 分割された2つの領域$L1$, $L2$について同様の処理を繰り返す。
4. 全ての分割がk個未満の部分領域を生成するとき終了



$$\forall i : k \leq |L_i| \leq 2k - 1$$

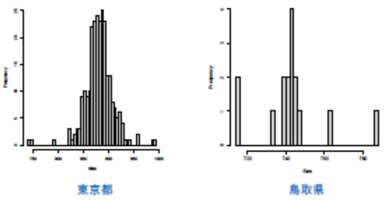
19

等価クラスのサイズの頻度分布（k=500の場合）



20

等価クラスのサイズの頻度分布（k=500の場合）



21

地域レベルの一般化による粒度調整アルゴリズム

- 提案手法との比較のため、地域レベルの一般化による匿名化アルゴリズムと比較
- アルゴリズムの概略
  - 都道府県→市区町村→丁目とk値を満足する場合は一つ下の詳細レベルに進む
  - k値を満足しなくなったレベルで同じレベルのレコードをランダムにマージして、k個以上の等価クラスを生成

22

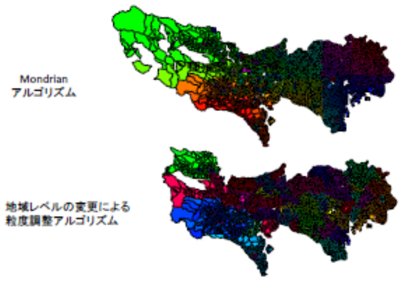
有用性評価

- 等価クラスのサイズの分散，Discernibility指標 については、GPS座標に基づく分割のほうが優位性を示している

Prefixname	Groups Count	Max	Min	Mean	Median	Standard Deviation	Standard Error	Discernibility Metric
divideRegion 500	4831	999	339	578.35	578	61.2886	0.8809	1398323316
divideRegion 1000	2831	1957	582	1149	1139	95.1861	2.1112	2699184756
divideRegion 2000	1431	3785	1577	2249	2278	117.9335	3.4998	5948853372
adjustGranularity 500	3933	1312	300	592	518	192.703	3.1197	1488892118
adjustGranularity 1000	1895	2794	1080	1233	1075	290.6573	6.665	3838382882
adjustGranularity 2000	961	4769	2081	2987	2454	592.4884	19.3941	6341883344

23

東京都の領域分割の可視化



グループに属する丁目が必ずしも隣接していない

24

まとめと今後の課題

- 医療情報に含まれる地域情報を準識別子としし、GPS座標による柔軟な領域分割を目指した
- GPS座標データとのリンケージには、欠損値処理等多くの現実的な問題が存在
- k-匿名化アルゴリズムとして、トップダウン方式で再帰的に領域分割をおこなうMondrianアルゴリズムをR言語で実装
- 地域レベルの一般化による匿名化処理と比較すると、等価クラスのサイズに関するばらつき、Discenability指標において改善がみられた
- 今後の課題としては、作成される等価クラスの隣接性の定量的評価を行う予定である

25

資料 3　地理情報が公表された場合の安全性についての評価

表 1-1. 市町村別がん患者数とがん患者が 1 人になる市町村数（5 歳階級/男女）

年齢階級	市町村数	市町村別患者数 (min-max)	がん患者が 1 人となる市町村数
0～4	574	1-39	313
5～9	455	1-10	308
10～14	509	1-8	316
15～19	632	1-14	333
20～24	823	1-25	342
25～29	1094	1-55	318
30～34	1350	1-103	306
35～39	1514	1-140	275
40～44	1603	1-256	217
45～49	1685	1-387	164
50～54	1733	1-466	135
55～59	1817	1-514	101
60～64	1844	1-598	56
65～69	1875	1-763	32
70～74	1869	1-1097	33
75～79	1876	1-1114	26
80～84	1870	1-887	17
85～89	1866	1-616	33
90～94	1829	1-317	102
95～99	1586	1-86	283
100～104	717	1-12	403
105～109	80	1-3	74

表 1-2. がん患者が 1 人になる市町村の割合（5 歳階級/男女）

年齢階級	1 人になる市町村数／市町村数（%）
70～74	0.9
75～79	1.4
80～84	1.7
65～69	1.8
85～89	1.8
60～64	3
55～59	5.6
90～94	5.6
50～54	7.8
45～49	9.7
40～44	13.5
35～39	17.8
95～99	18.2
30～34	22.7
25～29	29.1
20～24	41.6
0～4	52.7
15～19	54.5
10～14	56.2
5～9	62.1
100～104	67.7
105～109	92.5

表 2-1. 市町村別がん患者数とがん患者が 1 人になる市町村数（5 歳階級/男）

年齢階級	市町村数	市町村別患者数 (min-max)	がん患者が 1 人となる市町村数
0～4	397	1-21	270
5～9	297	1-5	228
10～14	343	1-6	243
15～19	404	1-5	276
20～24	512	1-10	325
25～29	623	1-12	336
30～34	789	1-17	359
35～39	999	1-35	345
40～44	1232	1-62	329
45～49	1391	1-107	292
50～54	1562	1-163	260
55～59	1708	1-237	177
60～64	1803	1-307	100
65～69	1855	1-478	52
70～74	1860	1-687	45
75～79	1860	1-681	48
80～84	1855	1-520	49
85～89	1836	1-332	89
90～94	1710	1-124	192
95～99	1150	1-28	429
100～104	258	1-5	211
105～109	7	1	7

表 2-2. がん患者が 1 人になる市町村の割合（5 歳階級/男）

年齢階級	1 人になる市町村数／市町村数（%）
80～84	2.4
75～79	2.6
70～74	2.6
85～89	2.8
65～69	4.8
60～64	5.5
90～94	10.4
55～59	11.2
50～54	16.6
45～49	21
40～44	26.7
35～39	34.5
95～99	37.3
30～34	45.5
25～29	53.9
20～24	63.5
100～104	68
15～19	68.3
0～4	70.8
10～14	76.8
5～9	81.8
105～109	100

表 3-1. 市町村別がん患者数とがん患者が 1 人になる市町村数（5 歳階級/女）

年齢階級	市町村数	市町村別患者数 (min-max)	がん患者が 1 人となる市町村数
0～4	366	1-18	254
5～9	246	1-5	194
10～14	278	1-4	216
15～19	433	1-12	285
20～24	669	1-19	334
25～29	1000	1-43	318
30～34	1271	1-88	315
35～39	1440	1-108	307
40～44	1519	1-194	205
45～49	1616	1-280	205
50～54	1645	1-303	195
55～59	1712	1-277	176
60～64	1752	1-291	148
65～69	1797	1-323	105
70～74	1794	1-410	97
75～79	1809	1-433	81
80～84	1821	1-367	78
85～89	1813	1-284	105
90～94	1748	1-193	157
95～99	1478	1-58	335
100～104	610	1-8	379
105～109	68	1-3	65

表 3-2. がん患者が 1 人になる市町村の割合（5 歳階級/女）

年齢階級	1 人になる市町村数／市町村数（%）
80～84	4.3
75～79	4.5
65～69	5.4
70～74	5.8
85～89	5.8
60～64	8.4
55～59	9
90～94	10.3
50～54	11.9
45～49	12.7
40～44	13.5
95～99	21.3
35～39	22.7
30～34	24.8
25～29	31.8
20～24	49.9
15～19	62.1
0～4	65.8
100～104	69.4
10～14	77.7
5～9	78.9
105～109	95.6

表 4-1. 市町村別がん患者数とがん患者が 1 人になる市町村数（10 歳階級/男女）

年齢階級	市町村数	市町村別患者数 (min-max)	がん患者が 1 人となる市町村数
0～9	758	1-49	361
10～19	820	1-21	350
20～29	1207	1-80	303
30～39	1614	1-243	219
40～49	1770	1-643	128
50～59	1847	1-980	59
60～69	1883	1-1361	23
70～79	1887	1-2145	13
80～89	1885	1-1503	12
90～99	1849	1-389	75
100～109	749	1-12	412

表 4-2. がん患者が 1 人になる市町村の割合（10 歳階級/男女）

年齢階級	1 人になる市町村数／市町村数（%）
80～89	0.6
70～79	0.7
60～69	1.2
50～59	3.2
90～99	4.1
40～49	7.2
30～39	13.6
20～29	25.1
10～19	42.7
0～9	47.6
100～109	55.0

表 5-1. 市町村別がん患者数とがん患者が 1 人になる市町村数（10 歳階級/男）

年齢階級	市町村数	市町村別患者数 (min-max)	がん患者が 1 人となる市町村数
0～9	559	1-26	338
10～19	596	1-9	345
20～29	790	1-18	334
30～39	1151	1-47	316
40～49	1521	1-169	247
50～59	1775	1-400	124
60～69	1873	1-751	38
70～79	1882	1-1334	21
80～89	1879	1-582	22
90～99	1736	1-146	166
100～109	267	1-5	214

表 5-2. がん患者が 1 人になる市町村の割合（10 歳階級/男）

年齢階級	1 人になる市町村数／市町村数（%）
70～79	1.1
80～89	1.2
60～69	2.0
50～59	7.0
90～99	9.6
40～49	16.2
30～39	27.5
20～29	42.3
10～19	57.9
0～9	60.5
100～109	80.1

表 6-1. 市町村別がん患者数とがん患者が 1 人になる市町村数（10 歳階級/女）

年齢階級	市町村数	市町村別患者数 (min-max)	がん患者が 1 人となる市町村数
0～9	495	1-23	304
10～19	562	1-16	321
20～29	1100	1-62	297
30～39	1557	1-196	244
40～49	1716	1-474	157
50～59	1790	1-580	105
60～69	1845	1-614	62
70～79	1847	1-811	44
80～89	1859	1-651	33
90～99	1796	1-243	136
100～109	642	1-8	394

表 6-2. がん患者が 1 人になる市町村の割合（10 歳階級/女）

年齢階級	1 人になる市町村数／市町村数（%）
80～89	1.8
70～79	2.4
60～69	3.4
50～59	5.9
90～99	7.6
40～49	9.1
30～39	15.7
20～29	27.0
10～19	57.1
100～109	61.4
0～9	61.4