

厚生労働科学研究費補助金  
循環器疾患・糖尿病等生活習慣病対策総合研究事業

**追跡終了後コホート研究を用いた  
共通化データベース基盤整備と  
その活用に関する研究**

平成 25 年度 総括・分担研究報告書

研究代表者 玉腰 暁子

平成 26 ( 2014 ) 年 3 月

**追跡終了後コホート研究を用いた  
共通化データベース基盤整備と  
その活用に関する研究**

平成 25 年度 総括・分担研究報告書

研究代表者 玉腰 暁子  
平成 26 ( 2014 ) 年 3 月

目 次

<b>I. 総括研究報告</b>	
追跡終了後コホート研究を用いた共通化データベース基盤整備とその活用に関する研究：進捗報告.....	1
玉腰暁子	
<b>II. 分担研究報告</b>	
米国におけるデータアーカイブの研究利用：現状と課題.....	5
大橋靖雄、祖父江友孝、他	
統計行政におけるデータ利用の動向.....	8
大橋靖雄、祖父江友孝、他	
データアーカイブ利用に際して必要と考えられる研究倫理教育.....	11
辻一郎、磯博康、他	
ヒトに由来するデータ活用と知的財産・法的課題.....	14
磯博康、辻一郎、他	
大規模コホートデータにおける一意性の検討.....	17
祖父江友孝	
Japan Arteriosclerosis Longitudinal Study (JALS)データにおける一意性の検討.....	39
大橋靖雄、原田亜紀子、他	
疫学研究のデータアーカイブ化の試み.....	追 1
辻一郎、祖父江友孝、他	
<b>III. 資料</b> .....	53

## 追跡終了後コホート研究を用いた共通化データベース基盤整備と その活用に関する研究：進捗報告

研究代表者 玉腰暁子（北海道大学大学院医学研究科・教授）

### 研究要旨

国内で実施され追跡を終了した複数のコホート研究情報を共通化し、その利用環境を整え、将来にわたって向後終了するコホート研究も組み入れ可能な体制を構築するために必要な事項を検討する。そのため、初年度は、米国のデータ利用環境、統計法に基づき実施された国内統計資料の利用のあり方、データ二次利用にまつわる倫理的課題と知財、研究倫理教育のあり方について、現状を確認し、課題を整理するとともに、各コホートが持つデータを公開した場合の一意性の問題について、実データを基に検討した。その結果、統計法の規定から、人口動態統計資料から得られた死因情報をコホート研究のアウトカムとして公開することはできないことが明らかとなった。そのため、コホート研究データをアーカイブ化し二次利用を進めるためには、死因情報のソースとなる人口動態統計調査の統計法内における位置づけを変更することが必要と考えられた。また、二次利用に際しては、対象者の同意の範囲への考慮とそれに伴う匿名化情報としての取り扱いはもちろん、二次利用者の責務を意識することが重要である。さらに、特に試料の利用にあたっては、知財に関し、関係者で事前に十分に検討し、ルールを明確にしておくことが必要と考えられた。なお、多くの項目を組み合わせれば一意性があるものとの前提で、二次利用対応を考えていくことが必要である。

### 分担研究者

磯博康（大阪大学大学院医学系研究科・教授）

大橋靖雄（東京大学大学院医学系研究科・教授）

祖父江友孝（大阪大学大学院医学系研究科・教授）

辻一郎（東北大学大学院医学系研究科・教授）

### A. 研究目的

本研究では、国内で実施され追跡を終了した複数のコホート研究情報を共通化し、その利用環境を整え、将来にわたって向後終了するコホート研究も組み入れ可能な体制を構築するために必要な事項を検討する。初年度は、米国のデータ利用環境、統計法に基づき実施された国内統計資料の利用のあり方、データ二次利用にまつわる倫理的課題と知財、研究倫理教育のあり方について、現状を確認し、課題を整理した。また、各コホートが持つデータを公

開した場合の一意性の問題について、実データを基に検討した。

### B. 研究方法

1. 各分野の専門家から現状を伺い、追跡を終了したコホート研究データを二次利用する際の課題を整理した。
2. 三府県コホートデータと JALS 死亡者データを用い、それぞれどのような頻度で一意性が見られるかを検討した。

### C. 研究結果

1. 資料・データベースの二次利用の現状と課題
  - 1-1. 米国における二次データ利用の現状  
米国では、電子カルテデータ、レセプトデータ、患者レジストリデータのみならず、コホート研究デ

ータについても公開が始まっている。これらのデータを用いることで、コホート研究を1から始めるのに比べ少ない労力でデータセットを作ることが可能である。また、NHLBI から研究費を受けた行われたRCTとコホート研究を含めた全てのデータを最大利用することを意図して、Biologic Specimen and Data Repository Information Coordinating Centerを事務局としたデータデポジトリが行われている。データ提供にあたっては、インフォームド・コンセントに沿った形にすること、個人同定情報は削除すること、地域情報は外すことなどがガイドラインで規定されている。このデポジトリでは、データあるいは生体試料二次利用希望者はHP上で、各研究の詳細を確認し、自身の用いたい研究に対しリクエストをすることができるが、その際、倫理審査を受けておくなどの規則も定められている。このシステムにより、すでに他国からの申請も含め多くの研究が実施されている。

#### 1-2. 日本における公的統計の利用

公的統計は、その体系的かつ効率的な整備及びその有用性の確保を図ることを目的とする統計法で規定されている。2007年の全部改正にあたっては、社会の情報基盤としての統計というコンセプト、そしてそれを実践するための統計データの利用促進、特に二次利用に積極的に取り組むことが盛り込まれた。そのため現在では、オーダーメイド集計や匿名データの利用が推奨されている。なお、ほぼ全てのコホート研究で死因情報は、人口動態統計調査の利用手続きを経て入手されていることから、この情報を更に別の二次利用(提供)に供することは現行の統計法の下では不可能である。

#### 1-3. 試料・情報の二次利用に関する説明と同意

新しく開始される研究では、二次利用まで見越した同意を取得することが望ましい。当初の研究会市時点では具体的な二次利用の詳細は決まっていないうことも多いが、広範同意で二次利用に関する同意も得、実際に二次利用を進める際には倫理審査委員会など第三者の意見も聞いたうえで実施にあたってはオプト・アウト方式(研究に関する情報を公開し、拒否権を保障する)を採用することが推奨され

る。それに対し、既に進んでいるコホート研究では、そのような二次利用を想定していないことが多く、一次同意の範囲では二次利用をカバーできていない。そのような試料・情報であっても有効に研究に活用するためには、対象地域や対象者に対する情報公開を多面的かつ十分に行ったうえで拒否権を保障することが重要であると考えられる。また、二次利用者も研究にあたっては情報公開をしっかりと行うとともに、試料・情報提供者に対し、研究成果に関する報告や有用な個別結果の返却など、できるだけ利益を還元するよう務めることも必要であろう。

#### 1-4. 試料・情報二次利用に必要な研究倫理教育

二次利用に供される試料・情報の特徴を考えると、データアーカイブを用いて研究を行う二次利用者に対しては、以下の点を目指した研究倫理教育が必要と考えられた。

- 一次収集者の苦勞を知る
- 試料・情報の背後には一人ひとりの生身の人間がいることを知る
- ヒト試料研究一般に求められる倫理性を理解する
  - ヒト試料の地位
  - 同意取得の原則
  - 同意原則の緩和が許容されるための条件と発生責務
  - 個人情報の安全管理措置
- 各ステークホルダーに課せられる倫理的責務を知る
- 知財・オーサーシップ等についての適切な知識を持つ
- データ不正等の罪深さを知る
- 情報倫理について理解する

#### 1-5. 試料・情報二次利用に関連する知財

ヒトに由来する試料は、細胞や血液といった形でそのヒトの中にあるときは、当然そのヒトに帰属する。しかし、研究利用に関する説明が行われ、同意を得た上で採取された細胞や血液は、インフォームド・コンセントで示された条件で使用するという約束の下でドナーが譲渡したものと考えることができる。また、その試料・情報を用いて研究が行われ

た場合、研究によって試料・情報に付加価値がつく。したがって、いったん研究に提供された試料・情報を用いて行われた研究の成果として知的財産権が発生した場合、その権利は試料・情報提供者ではなく、研究者(研究機関)に帰属すると考えるのが適切である(対象者への説明時点でももちろん述べておくべき内容である)。バンクや共同研究の形で提供される試料・情報を用いた二次利用で開発された研究成果に関する特許権の分配方法については、事前に取り決めを交わしておくのが望ましい。一次情報収集者の努力、バンク運用者の作業を考えると、そこにどう報いるかを二次利用者は常に意識することも重要であろう。このことは特許権の配分のみならず、論文の共著者に加えるかどうか、謝辞に述べるかどうかも含めて求められる。

## 2. 一意性の検討

### 2-1. 三府県コホート対象者に関する検討

三府県コホート対象者10万人強について204変数を分析対象として検討したところ、変数を1つずつ個別に見た場合の一意性は小さかったが、全変数を組み合わせた場合、一意であるレコード数は対象者の約99.98%であった。一変数で見た場合、一意性には、10年観察終了日のように、分類数が大きいことでそれぞれに振り分けられる個体数が少なくなるため生じる一意と、出産人数において出産人数が20人というように、疫学的にまれな属性の個体が存在したために生じる一意の大きく2パターンが存在した。

### 2-2. JALS 対象者中の死亡者に関する検討

複数のコホートが共同で実施しているJALSのデータを用いて、死亡者約5000人の性、ベースライン時年齢、死亡時年齢、地域、死亡年データによる個人特定可能性につき検討した。性、ベースライン時年齢単独での特定性は低かったが、死亡時年齢が40歳代もしくは90歳代後半である場合や小規模コホートの対象者で特定可能性が高くなった。さらに死亡年月日を月までで丸めた場合でも個人が一意に特定される場合が観察された。また、これら変数を2つ以上組み合わせた場合には、一意に特定されるものの数が上昇した。

## D. 考察

### 1. 試料・情報二次利用の現状と課題

日本よりデータベースの二次利用が進んでいる米国において、研究に用いられているデータベースの種類、活用事例とその成果、二次利用を進めるにあたっての留意点等について確認した。日本では、二次データの公開システムとしてバイオサイエンスデータベースセンター(NBDC)が立ち上がったばかりであり、仕組みそのものに関する議論が十分に行われてはいない。今後、日本においても公的研究費を受けた研究を適切に二次利用することが求められるが、そのためには、個人情報保護と研究活用とのバランス、事務手続きの標準化・単純化と必要経費、データの適正使用と質保障のためのサポート、共通化によるデータマニピュレーション、情報のロスとデータ容易使用のバランス、共通化プロセスの透明化と公正なシェアの仕組み等につき、議論を重ねていくことが必要と考えられた。

国内の統計行政に関しては、統計法で規定されている。改正統計法は、社会基盤としての統計をコンセプトに、公的統計の体系的かつ効率的な整備及びその有用性の確保を図ることが目的とされている。公的統計の利用に関しても範囲が拡大され、秘匿しない調査票情報の提供、秘匿処理した調査票情報の利用(匿名データの利用)、行政機関に特別集計を依頼(オーダーメイド集計)が可能である。なお、提供された調査票情報や匿名データなどを統計の作成に関連する目的以外に利用・提供した者や、守秘義務規定に違反した者に対して、罰則が定められている。したがって、コホート研究データをアーカイブ化し二次利用を進めるためには、アウトカムに用いている死因情報のソースとなる人口動態統計調査の統計法内における位置づけを変更することが必要と考えられた。

研究倫理の観点からは、試料・情報の二次利用に際しては、対象者の同意の範囲への考慮とそれに伴う匿名化情報としての取り扱いはもちろん、二次利用者の責務を意識することが重要である。したがって、アーカイブやバンクによるデータ二次利用を行う研究者に対する研究倫理教育では、通常教育研

修内容に加え、二次利用に関する同意のあり方、オープンシップや知財に関する内容を含めることが必要と考えられた。また、アーカイブやバンクによるデータ二次利用を進めるためには、その後が生じ得る知的財産検討に関し、事前に関係者間で十分に検討することが必要と考えられた。

## 2. 一意性の検討

一変数で見た場合、一意性には、分類数が大きいことでそれぞれに振り分けられる個体数が少なくなるため生じる一意と、疫学的にまれな属性の個体が存在したために生じる一意の大きく2パターンが考えられた。前者に対しては例えば日付データを月までにするなどにより分類数を減らすことで一意性を減少させることが可能であり、後者に対しては一定値以上（以下）については直接表示せず、無限までの片側区間で表示するといった方法により一意性の減少が図られる。しかし、すべての変数を組み合わせた場合には全レコード数の約99.98%が一意であったが、変数が多くなる（質問項目が多い）ことによる一意性は容易に避けられるものではない。また、10万人規模のデータではまるめ処理などによりある程度の一意性の減少がみられるが、規模が小さくなると一意性が上がる可能性も高い。したがって、多くの項目の組み合わせを行う場合には、一意性があるものとの前提で対応を考える必要がある。

また、複数の比較的小規模なコホートをあわせて実施されているコホート研究における死亡者データで検討したところ、地域の情報が特に個人の特定に繋がりがやすいこと、死亡時年齢が若年もしくは高齢である場合に一意性が高じることが明らかとなった。また、死亡年月日については、月までに丸めた場合でも個人が一意に特定される場合があることから、可能な限り死亡年までの情報とすることが求められる。さらに死亡者につき用いる情報（性、ベースライン時年齢、死亡時年齢、地域、死亡年）を一変数ではなく二変数以上とした場合には、組み合わせによって一意性が上昇した。したがって、分類を粗くするなどの対応が必要と考えられた。

## E. 結論

米国のデータ利用環境、統計法に基づき実施された国内統計資料の利用のあり方、データ二次利用にまつわる倫理的課題と知財、研究倫理教育のあり方について、現状を確認し、課題を整理するとともに、各コホートが持つデータを公開した場合の一意性の問題について、実データを基に検討した。その結果、統計法の規定から、人口動態統計資料から得られた死因情報をコホート研究のアウトカムとして公開することはできないことが明らかとなった。そのため、コホート研究データをアーカイブ化し二次利用を進めるためには、死因情報のソースとなる人口動態統計調査の統計法内における位置づけを変更することが必要と考えられた。また、二次利用に際しては、対象者の同意の範囲への考慮とそれに伴う匿名化情報としての取り扱いはもちろん、二次利用者の責務を意識することが重要である。さらに、特に試料の利用にあたっては、知財に関し、関係者で事前に十分に検討し、ルールを明確にしておくことが必要と考えられた。なお、多くの項目を組み合わせれば一意性があるものとの前提で、二次利用対応を考えていくことが必要である。

## F. 健康危機情報

なし

## G. 研究発表

1. 論文発表  
なし
2. 学会発表  
なし

## H. 知的財産権の出願・登録状況（予定を含む）

1. 特許取得  
なし
2. 実用新案登録  
なし
3. その他  
なし

## 米国におけるデータアーカイブの研究利用:現状と課題

研究分担者	大橋靖雄	東京大学大学院医学系研究科
研究分担者	祖父江友孝	大阪大学大学院医学系研究科
研究代表者	玉腰暁子	北海道大学大学院医学研究科
研究協力者	瀬戸口聡子	Duke 大学医学部

### 研究要旨

米国で研究に用いられているデータベースの種類、活用事例とその成果、二次利用を進めるにあたっての留意点等について確認した。今後、日本においても公的研究費を受けた研究を適切に二次利用することが求められているが、そのためには、個人情報保護と研究活用とのバランス、事務手続きの標準化・単純化と必要経費、データの適正使用と質保障のためのサポート、共通化によるデータマニピュレーション、情報のロスとデータ容易使用のバランス、共通化プロセスの透明化と公正なシェアの仕組み等につき、議論を重ねていくことが必要と考えられた。

### A. 目的

国内ではまだ十分に活用が進んでいないデータベースの研究利用に関し、米国の事例と留意点を確認し、今後の方向性を考える一助とする。

### B. 方法

実際に米国でデータベースを用いた研究に従事されている Duke 大学瀬戸口聡子准教授より、情報を得た。

### C. 結果

[米国で比較的高頻度に研究活用されているデータベース]

電子カルテデータ: 病院毎に異なるシステムを導入しており、それぞれ研究用 DB 化を試みているが、統一した全国的なデータベースは存在しない

- EPIC - 近年、このシステムを導入する病院が増加

クレームデータ(レセプトデータ): 保険システム毎に存在

- Medicaid - 貧困層
- Medicare - 高齢者・一部の疾患

患者レジストリデータベース: 主に学会が作成登録

- National Cardiovascular Data Registry - American College of Cardiology、主に手技(ステント、PCI など) 毎に患者を登録
- STS National Database - Society of Thoracic Surgeons、胸部外科の各領域(先天性心臓外科、成人心臓病手術、呼吸器手術)手術の登録
- SEER (Surveillance Epidemiology and End Results) Program - National Cancer Act に基づく州単位のがん登録のうち精度のよいもの

コホート研究データベース:二次利用のために公開されているものもあるが、まだそれほど広がっていない。現在、National Heart, Lung, and Blood Institute (NHLBI)が研究費を受けて作ったデータに関しては3年経過後には提出するよう求めている

- ARIC
- Framingham Study
- Women's Health Initiative

#### [データベースの結合利用事例]

米国では、データベース同士の結合も進められており、例えば ARIC、Framingham Study は Medicare Data とリンクされている。

また、データベース内にある ID、あるいは生年月日や病院 ID、入院日を利用して、患者や医師情報を結合することが可能であり、複数のデータベースを用いた研究を行うことができる。治療法による予後を比較、(埋め込み型除細動器と頸動脈ステント術)する研究では、患者レジストリデータベース (Society for Vascular Surgery's Vascular Registry (SVS-VR)と Carotid Artery Revascularization and Endarterectomy (CARE)) に Medicare Data、さらに有料の病院情報、医師情報 (American Hospital Association)などを結合した。このような方法により、コホート研究を1から始めるのに比べ、少ない労力でデータセットを作ることが可能であるが、データ入手の事務作業等で手元にデータが来るまでに2年かかることもある。なお、患者個人のIDを用いなくても、生年月日、医療手技、入院日とプロバイダーIDを利用すれば、95%以上が正しく突合された。また、米国では保険に加入する際サインする同意書に、データの研究利用の項目があるため、保険者がデータを研究利用するために改めて同意を得ることはされず、またIDを削除したデータが販売も行われているのが現状である(ただし、州により対応が異なる場合がある)。

#### [データ二次利用システム事例]

<https://biolincc.nhlbi.nih.gov/home/>

米国では、NHLBI から研究費を受けた行われた RCT とコホート研究を含めた全てのデータを最大利用することを意図して、Biologic Specimen and Data Repository Information Coordinating Center を事務局としたデータデポジトリが行われている。

データ提供のタイミングは、RCT であれば最後の患者登録から3年後、または主目的の論文が公表されてから2年後、観察型の疫学研究では追跡終了の3年後、または付随的な研究から2年後である。その場合、インフォームド・コンセントに沿った形にすること、個人同定情報は削除すること、地域情報は外すことなどがガイドラインで規定されている。データあるいは生体試料二次利用希望者は、HP 上で各研究の詳細を確認し、自身の用いたい研究に対しリクエストをすることができるが、その際、倫理審査を受けておくなどの規則も定められている。

2014年2月時点で確認したところ、134の研究(臨床研究91件、疫学研究42件、両者1件)が登録されていた。なお、2000~10年(2010年時点では登録は76研究)の間に、498のリクエストが認められ、うち74件(15%)が他国の研究所からであり、224論文が公表されたとのことである。

#### [データベース標準化の方向性]

米国では、各所に散らばるデータを1箇所に集めるのではなく、それぞれの場所に置いたまま共同研究ができるようデータフォーマットを統一する distributed network model という試みが始まっている。例えば、OMOP(Observational Medical Outcomes Partnership)は政府と民間の共同によるクレームデータや電子カルテをつなぐデータベースネットワーク構築とリサーチ手法の研究の試みである。その際、項目名、変数の与え方なども全て統一化する(コモンデータモデル)ことにより、複数のデータベースで同時に同じ解析を行うことを目的として、現在米国内の10データソース、1億3000万人をカバーする規模で進められている。また、FDAのmini-sentinelは特に薬の市販後調査を年頭に、同様の試みをしている。ここでは、18の保険会

社データベースを用いており、1億3000万人がカバーされている。

#### D. 考察

米国では、NHLBI から研究費を受けて実施された臨床研究、疫学研究は研究終了後一定期間が経過した後は、公開される仕組みが整えられている。しかし、日本ではまだその議論が始まったばかりであり、バイオサイエンスデータベースセンター(NBDC)が立ち上がってはいるものの、仕組みそのものに関する議論が十分に行われてはいない。また既存データベース同士を個人単位で連結することで、データ価値が倍増することも考えられるが、そのための仕組みは整っていない。

今後、観察が終了したコホート研究の二次利用を適切に進めるためには、以下の点に関し、議論を重ね、体制を構築していくことが重要と考えられる。

- 個人情報保護と研究活用のためのバランス
- 事務手続きの標準化・単純化と必要経費
- データの適正使用と質保障のためのサポート
- 共通化によるデータマニピュレーション、情報のロスとデータ使用簡易性のバランス
- 共通化プロセスの透明化と公正なシェアの仕組み

#### E. 結論

米国でのデータベース活用とその成果、留意点等について確認した。それらを参考に、今後、日本においても公的研究費を受けた研究を適切に二次利用することが求められているが、そのために検討すべき課題を列記した。

#### F. 研究発表

1. 論文発表
  2. 学会発表
- いずれもなし

#### G. 知的財産権の出願・登録状況

(予定を含む。)

1. 特許取得
  2. 実用新案登録
  3. その他
- いずれもなし

## 統計行政におけるデータ利用の動向

研究分担者	大橋靖雄	東京大学大学院医学系研究科
研究分担者	祖父江友孝	大阪大学大学院医学系研究科
研究代表者	玉腰暁子	北海道大学大学院医学研究科

### 研究要旨

国内の統計行政に関し、特に改正統計法を中心に、情報を得た。改正統計法では、社会基盤としての統計をコンセプトに、公的統計の体系的かつ効率的な整備及びその有用性の確保を図ることが目的とされている。公的統計の利用に関しても範囲が拡大され、秘匿しない調査票情報の提供、秘匿処理した調査票情報の利用(匿名データの利用)、行政機関に特別集計を依頼(オーダーメイド集計)が可能である。なお、提供された調査票情報や匿名データなどを統計の作成に関連する目的以外に利用・提供した者や、守秘義務規定に違反した者に対しては、罰則が定められている。したがって、コホート研究データをアーカイブ化し利用を進めるためには、アウトカムに用いている死因情報のソースとなる人口動態統計調査の統計法内における位置づけを変更することが必要と考えられた。

### A. 目的

追跡が終了したコホート研究をアーカイブ化することを念頭に、現在の統計行政におけるデータ利用の動向について確認する。

有している。政府統計は約 500 あり、実際に毎年実施されている統計は 300 弱である。独立行政法人統計センターは、各府省が実施する重要な統計を作成、あるいは必要な統計技術の研究等を行う役割を担っている。

### B. 方法

総務省統計局で統計情報作成に携わっている上田聖氏より、情報を得た。

[匿名データの作成と利用]

元来、統計は行政のため、という位置づけであった。そのため、2007 年の改正統計法が公布されるまでの統計法では、第 15 条「何人も、指定統計を作成するために集められた調査票を、統計上の目的以外に使用してはならない」とされ、データ利用を強く制限する方針であった。第 2 項に「前項の規定は、総務大臣の承認を得て使用の目的を公示したのものについては、これを適用しない」との例外規定が置かれていたものの、

### C. 結果

[統計行政]

日本では、各府省がそれぞれ必要な統計を作成する分散型のシステムを採用している。その中で総務省は、国勢調査や経済センサスという全数調査等を所管しているほか、各府省の調査に対する許認可権を

統計利用の運用上は第1項が中心となっていた。

1980年代に入ると、欧米では個人情報秘匿した匿名データを作成するようになり、その利用が活発になる。さらにパブリックユースファイルといわれるネット上でダウンロードできる仕組みも作られるようになってきた。それに対し、日本は調査票利用のハードルを下げてきたが、十分な対応はできなかった。そのため、いくつかの調査について、匿名データを作成し、提供することを開始した。

#### [改正統計法]

統計法は、公的統計の体系的かつ効率的な整備及びその有用性の確保を図ることを目的としている。2007年の全部改正にあたっては、社会の情報基盤としての統計というコンセプト、そしてそれを実践するための統計データの利用促進、特に二次利用に積極的に取り組むことが盛り込まれた。

#### [統計データの利用促進]

改正統計法では、従前の運用を引き継ぎ以下の1.が規定されるとともに、調査票情報の利用範囲が以下の2.及び3.として拡大された。

1. 秘匿しない調査票情報の提供(第33条): 行政機関、地方公共団体、独立行政法人、と共同研究を行う者、もしくは委託研究を行う者、公的な競争研究資金を受けている者(省令第9条)
2. 秘匿処理した調査票情報の利用(匿名データの利用)手数料あり(第35条・36条): 学術研究を行う者、高等教育を行う者
3. 行政機関に特別集計を依頼(オーダーメイド集計)手数料あり(第34条): 学術研究を行う者、高等教育を行う者

提供された調査票情報や匿名データなどを統計の作成に関連する目的以外に利用・提供した者や、守秘義務規定に違反した者に対して、罰則が定められている(第42条・43条、罰則規定は第7章)。また現在までに、6つの匿名データ化された統計調査(例:

全国消費実態調査、住宅・土地統計調査等)が利用可能となっており、積極的な研究利用が推奨されている。なお、平成24年7月の資料(調査票情報の二次的利用の状況について)によれば、平成21~23年度中のオーダーメイド集計の利用件数は、10調査26件である。

これらとは別に、調査票情報を高次元クロス集計した統計(擬似マイクロデータ、1セル内の客対数は3以上)が作成され、提供されている。これは、演習用として多数ダウンロードされているほか、オーダーメイド集計を依頼する前段階のデータチェック用、SAS社のデータ分析コンテストの対象データとしても用いられている。

#### [次期統計行政の動向]

現在、平成26~30年度の計画の策定作業中である。その中で、調査票情報の提供に関して、セキュリティの観点からオンサイト利用やプログラム送付型集計・分析に段階的に移行することが検討されている。また、オーダーメイド集計については、利用条件が緩和される方向である。

#### [統計データの利用促進]

ビッグデータ時代に入っていることもあり、統計センターでは、統計API(application programming interface)、統計GIS、データキューブ、モバイル空間統計に関する取り組みが始められている。前2者は、世界最先端IT国家創造宣言(平成25年6月14日閣議決定)で取り上げられたものである。

## D. 考察

統計行政の歴史的背景と現状、今後の見通しについて、概観した。社会基盤としての統計というコンセプトから統計法が全部改正され、さらにビッグデータの時代に入っていることから、調査票情報の利用範囲は広がりつつある。

一方で、観察が終了したコホート研究データには、重要なアウトカムとして、死亡に関する情報が含まれ

ている。公的研究費により実施されるライフサイエンス分野の研究では、現在、論文発表等で公表された成果に関わる生データの複製物、又は構築した公開用データベースの複製物を、バイオサイエンスデータベースセンターに提供することが求められている。また、多額の費用と長期の追跡、そして多くの人々の協力の下で蓄積されてきたコホート研究データをアーカイブ化し、追跡調査/研究終了後も適切に利用することは重要である。しかし、ほぼ全てのコホート研究では、特に死因情報は人口動態統計調査の利用手続きを経て入手されていることから、この情報を更に別の二次利用(提供)に供することは現行の統計法の下では不可能である。したがって、追跡が終了したコホート研究のデータアーカイブ化を進めるためには、戸籍法に基づき全国民対象に行われている死亡、出生等の届けから作成される人口動態統計調査の統計法制上の位置づけを変更しない限り対応できないことが判明した。

## **E. 結論**

統計行政に関し、特に改正統計法を中心に、情報を得た。コホート研究データをアーカイブ化し利用を進めるためには、アウトカムに用いている死因情報のソースとなる人口動態統計調査の統計法制上の位置づけを変更することが必要と考えられた。

## **F. 研究発表**

1. 論文発表
  2. 学会発表
- いずれもなし

## **G. 知的財産権の出願・登録状況**

(予定を含む。)

1. 特許取得
  2. 実用新案登録
  3. その他
- いずれもなし

## データアーカイブ利用に際して必要と考えられる研究倫理教育

研究分担者	辻一郎	東北大学大学院医学系研究科
研究分担者	磯博康	大阪大学大学院医学系研究科
研究分担者	大橋靖雄	東京大学大学院医学系研究科
研究代表者	玉腰暁子	北海道大学大学院医学研究科
研究協力者	松井健志	国立循環器病研究センター

### 研究要旨

データ二次利用に際しての研究倫理の課題、ならびに研究倫理教育の現状に関し、情報を得た。試料・情報の二次利用に際しては、対象者の同意の範囲への考慮とそれに伴う匿名化情報としての取り扱いはもちろん、二次利用者の責務を意識することが重要である。したがって、アーカイブやバンクによるデータ二次利用を行う研究者に対する研究倫理教育では、通常の教育研修内容に加え、二次利用に関する同意のあり方、オーサiershipや知財に関する内容を含めることが必要と考えられた。

### A. 目的

疫学研究で収集された個人単位のデータをアーカイブ化し、研究者の二次利用に供する場合に必要と考えられる研究倫理教育について検討し、今後の体制構築の一助とする。

### B. 方法

研究倫理の第一人者であり、その教育にも通じている松井健志氏より情報を得た。

### C. 結果

[データ二次利用に伴う倫理課題]

同意の範囲と匿名化

疫学研究で収集される試料・情報は対象者個人から生じる。研究参加にあたって対象者は、一次利用者(データ収集者)と同意文書を交わし、通

常は、この一次同意の内容を元に研究が実施される。そして、アーカイブ化やバンクはこれら一次的に集められた情報の二次利用と位置づけることができる。その際、当初から二次利用まで見越した同意を取得でいていけばよいが、多くは具体的な研究の詳細は決まっておらず、そこまでの対応は行われていない。

プライバシーを保護する方法として、同意と匿名化を挙げることができる。上述したように多くの場合、二次利用に関する同意まで一次利用時に得ておくことは困難である。そこで、EU 指令 95/46 号「個人データ処理にかかわる個人の保護及び当該データの自由な移動に関する欧州議会及び理事会の指令」では、個人情報の守秘について適切な安全管理措置をしていること、本人の同意を得ることが合理的に実現困難で、正当かつ十分な公共

の利益の見込みがある、という3条件全てを満たす場合は、提供者の同意を得ることなく情報にアクセスすることが許容された。その際にはリスクの低減が求められることから、プライバシー侵害が起りにくいよう、連結可能匿名化であれば対応表を第三者管理にする、連結不可能匿名化する、といった対応が生じてきた。しかし、このような匿名化による対応が進むと、そもそも提供者の同意そのものの存在意義が揺らぐ、という大きな問題が生じてくる。

そこで、第三の方策として、当初の同意から、二次利用に関して、広範同意を得ておくことが考えられる。広範同意とは、例えば「がん・脳卒中などの生活習慣病に関する医学研究に使うてよい」など、不確定ではあるもの用途・範囲が幾分限定されイメージしやすくなっているものである(これに対し、「将来の医学研究に使うてよい」など研究の範囲がまったく特定されていないものは包括同意と呼ばれている)。

したがって、新しく開始されるコホート研究では、当初より広範同意で二次利用に関する同意も得、実際に二次利用を進める際には倫理審査委員会など第三者の意見も聞いたうえで、実施にあたってはオプト・アウト方式(研究に関する情報を公開し、拒否権を保障する)を採用することが推奨される。それに対し、既に進んでいるコホート研究では、当初二次利用を想定していないことが多く、一次同意では二次利用をカバーできていない。そのような試料・情報であっても有効に活用するためには、対象地域や対象者に対する情報公開を多面的かつ十分に行ったうえで拒否権を保障することが重要であると考えられる。この際、「同意」の意義を考えると、可能な限り、当初包括同意であっても、将来的な利用に関し何らかの同意が得られている対象者に限定するなどの対応が求められる。また、二次利用者も情報公開をしっかり行うとともに、試料・情報提供者に対し、研究成果に関する報告や有用な個別結果の返却など、できる

だけ利益を還元するよう務めることも必要であろう。

#### 二次利用者の責務

二次利用の際は、一次利用と異なり、研究者と対象者との関係性が必然的に薄くなる。したがって、利用している試料・情報は個人に由来する尊厳あるものであることに対し、敬意を払う気持ちを常に持つよう心がけることが重要である。また、一次利用者あるいはバンク等に対しては、得られた研究成果をフィードバックする、出所を明確化する、場合によっては論文著者に加えるなどの配慮も求められる。

#### [ データアーカイブ利用に際して必要な研究倫理教育 ]

そこで、データアーカイブを用いて研究を行う二次利用者に対しては、以下の点を目指した研究倫理教育が必要と考えられる。

- 一次収集者の苦勞を知る
- 試料・情報の背後には一人ひとりの生身の人間がいることを知る
- ヒト試料研究一般に求められる倫理性を理解する
  - ヒト試料の地位
  - 同意取得の原則
  - 同意原則の緩和が許容されるための条件と発生責務
  - 個人情報の安全管理措置
- 各ステークホルダーに課せられる倫理的責務を知る
- 知財・オーサーシップ等についての適切な知識を持つ
- データ不正等の罪深さを知る
- 情報倫理について理解する

今までのところ、研究倫理教育は主に臨床試験を対象として行われてきている。E-Learning教材では、国外では、NIH e-learning、OHRP e-learning for Assurance、FHI e-learning、国内では、ICR

## 必要となる研究倫理教育カリキュラムは？

- 一次収集サイトでのリクルート実地研修
- 講義・演習
  - 研究倫理原論
    - 研究不正（データ不正等）
    - 研究倫理の原則
  - ヒト資料研究・バイオバンクの倫理（概論）
    - ヒト資料研究における同意（各論）
    - ヒト資料研究と参加者への利益還元（各論）
    - アーカイブ化資料の二次利用に係る倫理（各論）
  - ヒト資料研究に係る規制・法令
    - 医学研究における個人情報保護（各論）
  - Authorship・知財とその配分：制度・契約等
  - 情報倫理 (cf. 「情報倫理の構築(FINE)」プロジェクト1996-2003: 京大、広大、平賀大:  
<http://www.ling.hus.kyoto-u.ac.jp/> [http://www.jpia.go.jp/5\\_dff/000scad/f/96000704.pdf](http://www.jpia.go.jp/5_dff/000scad/f/96000704.pdf))

臨床研究入門、CITI Japan、CRT-webなどをあげることができる。また、研修・セミナー型のものもいくつか行われている。しかし、機会はそれほど多くないのが現状である。それらのカリキュラム内容を参考に考えると、今後、データアーカイブ化を進めていくにあたり、二次利用者に対し新たに研究倫理教育を行う場合には、図に示したような内容を盛り込むことが有用と考えられる。

### D. 考察

疫学研究で得られたデータをアーカイブ化し二次利用体制を整備する場合に、二次利用者に必要な研究倫理教育内容につき、その課題を踏まえ検討した。

一次利用と異なり、二次利用に関しては研究の全体像を示しての同意は得られていないことが多い。そのような試料・情報であっても、時間と労力を費やし集められたデータであることから、有効に活用することが望まれるが、どのような対応方法が適切かについては、十分な議論が必要である。また、二次利用者は、データ提供者、一次利用者等に対し敬意を払って、研究を実施することが求められる。したがって、研究倫理教育内容には、通常研究者に求められるものに加え、データアーカイブやバンクに特徴的な同意、オーサiershipや知財に関する内容を含めることが必要と考えられた。

### E. 結論

データ二次利用に際しての研究倫理の課題、ならびに研究倫理教育の現状に関し、情報を得た。アーカイブやバンクによるデータ二次利用を行う研究者に対する研究倫理教育では、通常教育研修内容に加え、二次利用に関する同意のあり方、オーサiershipや知財に関する内容を含めることが必要と考えられた。

### F. 研究発表

1. 論文発表
  2. 学会発表
- いずれもなし

### G. 知的財産権の出願・登録状況

(予定を含む。)

1. 特許取得
  2. 実用新案登録
  3. その他
- いずれもなし

## ヒトに由来するデータ活用と知的財産・法的課題

研究分担者	磯博康	大阪大学大学院医学系研究科
研究分担者	辻一郎	東北大学大学院医学系研究科
研究分担者	大橋靖雄	東京大学大学院医学系研究科
研究代表者	玉腰暁子	北海道大学大学院医学研究科
研究協力者	隅藏康一	政策研究大学院大学

### 研究要旨

試料・情報二次利用に際しての知的財産権・法的な課題、ならびに遺伝子特許の現状に関し、情報を得た。アーカイブやバンクによるデータ二次利用を進めるためには、その後生じ得る知的財産検討に関し、事前に検討しておくことが必要と考えられた。

### A. 目的

疫学研究で収集された試料・情報を二次的に利用する場合の知的財産権・法的課題について整理し、今後の体制構築の一助とする。

### B. 方法

生命科学分野の知的財産権等に関する調査研究を実施されている政策研究大学院大学の隅藏康一氏より情報を得た。

### C. 結果

[知的財産権(特許権)の意義]

研究成果を実用化し社会に貢献することはアカデミアの使命の一つと認識されているが、近年特にその傾向が強まっている。研究成果を実用化するためには、アカデミアと企業が協力し合うことが必要であり、研究・開発のうちの開発段階は、最終製品を作り出す機能を持っていないアカデミアが単独で行うより、企業が市場動向を鑑みながら資金を投入して実施する

方が効率的である。そのように開発に企業が関与する場合にアカデミアの研究成果が特許化されていれば、その特許の使用許諾を受けて開発を行う企業を一社に絞ることができ、企業にとっては製品化後に市場を独占することが期待できることから、資金投入をしやすくする。

また、特許権を取得することにより、目的別に許諾条件を変えることが可能となる他、同一分野の研究推進に必要な特許権を押さえている機関と交渉する際に自身の保有する特許権とのクロスライセンスを持ちかけて交渉を有利に進めることもできる。したがって、特許制度は、アカデミアにおける基礎研究を円滑に進めることにも貢献している。このように、特許権は単に保有者に収入をもたらすのみならず、社会的な見地からも必要とされている。ただし、健康や医療に関する発明の場合、独占排他権である特許権と公共性のバランスが重要であることは言うまでもない。

[遺伝子特許]

ヒトゲノムと人権に関する世界宣言(UNESCO1997

年)では、第1条で「象徴的な意味において、ヒトゲノムは、人類の遺産である」と記されている。ヒトの遺伝子の解析により、遺伝子診断方法や医薬品の開発が進められ、人々の健康維持や新たな医療技術につなげられる。特許は発明に対して付与されるもので、その対象は、「物」と「方法」であり、自然界に天然に存在する物質はその対象とならない。ヒトの遺伝子解析に関しては、当初は分離・抽出された状態であれば天然物ではないというロジックで特許権が認められていた。しかし、ミリアッド社のBRCA遺伝子に対する特許をめぐる米国の裁判において、その特許性について議論が生じた。これまでは遺伝子特許により結果的にポリヌクレオチドにコードされた「情報」が保護されてきたが、今後は「物」としてcDNA(RNAから人工的に作成されたもののため、天然に存在するとはいえない)が特許取得対象となる。

#### [データ二次利用の意義]

通常、研究プロジェクトは一定期間が経過すると終了する。そのため、研究終了後は研究主体自体がなくなることになるが、それまでに集められた試料・情報を活用する方策を検討することは重要であり、現時点では、バイオバンクがその役を担うことが多いと考えられる。ただし、当初のインフォームド・コンセントの範囲の問題があるため、今後収集される試料・情報に関しては、バンクへの寄贈や二次利用を念頭に置いたインフォームド・コンセントを得ておくことが望ましいと考えられる。

#### [ヒト由来試料の帰属と特許権]

ヒトに由来する試料は、細胞や血液といった形でそのヒトの中にあるときは、当然そのヒトに帰属する。しかし、研究利用に関する説明が行われ、同意を得た上で採取された細胞や血液は、インフォームド・コンセントで示された条件で使用するという約束の下でドナーが譲渡したものと考えることができる。また、その試料・情報を用いて研究が行われた場合、研究によって試料・情報に付加価値がつく。したがって、いったん研

究に提供された試料・情報を用いて行われた研究の成果として知的財産権が発生した場合、その権利はサンプルの提供者ではなく、研究者(研究機関)に帰属すると考えるのが適切である(対象者への説明時点でももちろん述べておくべき内容である)。

バンクから提供される試料・情報を用いた二次利用で開発された研究成果に関する特許権の分配方法については、事前に取り決めを交わしておくのが望ましい。一次情報収集者の努力、バンク運用者の作業を考えると、そこにどう報いるかを二次利用者は常に意識することも重要であろう。このことは特許権の配分のみならず、論文の共著者に加えるかどうか、謝辞に述べるかどうかも含めて求められる。ちなみに、バイオバンクジャパンでは、収集機関、バイオバンク、研究機関の特許権持分比率は、4:3:3と覚書で定められている(ただし、変更は可能)。

#### D. 考察

今後のデータアーカイブ・バンク体制構築の一助とするため、ヒトに由来する試料・情報を用いて研究を実施する際の知的財産権のあり方、さらにそれを二次利用する場合の配分に関する現在の考え方を整理した。

ヒト由来試料から生じる特許権は、更なる研究開発のためにも重要である。その権利は通常、研究者(研究機関)に帰属し、試料提供者には認められていない。また、それらの情報がバンク等に集められたものを二次利用する場合には、事前に特許の配分や論文執筆につき契約書を交わしておくことが重要である。

#### E. 結論

試料・情報二次利用に際しての知的財産権・法的な課題、ならびに遺伝子特許の現状に関し、情報を得た。アーカイブやバンクによるデータ二次利用を進めるためには、その後生じ得る知的財産権等に関し、事前に検討しておくことが必要と考えられた。

## **F.研究発表**

1. 論文発表
2. 学会発表

いずれもなし

## **G.知的財産権の出願・登録状況**

(予定を含む。)

1. 特許取得
2. 実用新案登録
- 3.その他

いずれもなし

## 大規模コホートデータにおける一意性の検討

研究分担者 祖父江友孝 大阪大学大学院医学系研究科

### 研究要旨

個票データの開示を行う際には、一意性のあるデータは個人が同定される可能性があるの  
で、一意性のあるデータがどの程度存在するかを検討しておく必要がある。今回、三府県コ  
ホートデータにおいて、どのような頻度で一意性が見られるかを確認した。変数を 1 つずつ  
個別に見た場合の一意性は小さかったが、全変数を組み合わせた場合、一意であるレコー  
ド数は対象者の約 99.98%であった。複数の変数をそれぞれ組み合わせた場合の分類数 $K$   
とユニークセル数 $S_1$ のパターンから、分類数の増加に伴い一意であるレコード数は急  
増した。一意性は容易に避けられるものではなく、利用の際には一意性があるものと考えて  
対応することが必要と考えられた。

### A. 目的

三府県コホートデータについて、どのような頻  
度で一意性がみられるか検討する。

### B. 方法

三府県コホートデータを使用し 100,629 例全  
てについて検討を行う。各個人レコードは 226 変数  
からなるが、そのうち ID や数値化前データの変数、  
他と内容の重複する変数など 22 変数を除いた 204  
変数を分析対象とした(表 1)。

検討に際し変数をその内容の近いもの同士で組  
み合わせてカテゴリ化し 27 のカテゴリを作成し  
た。また、それらのカテゴリを内容から【個人特  
性】【追跡】【アンケート】の 3 グループに分けた  
(表 2)。

#### (1) 定義

対象の個体(本研究の場合は 100,629 例)が数

種類の変数の組み合わせに基づいて $K$ 個のセルに  
分けられたとき、1 つのセルに含まれる個体数が $i$   
のセル数を $S_i(i = 1, 2, \dots, N)$ とする。つまり、  
 $\sum S_i = N$ となる。今回注目するのは個体数が 1 のセ  
ルの数であるユニークセル数 $S_1$ である。なお、個  
体自体を呼ぶときには一意という単語を用いるが、  
セルに対してはユニークセルという単語を用いる。

#### (2) 検討内容

##### [検討 1]

204 変数それぞれ単変数についての、分類数 $K$ と  
ユニークセル数 $S_1$ を求めた。

##### [検討 2]

全体(204 変数すべてを組み合わせた場合)の  
分類数 $K$ とユニークセル数 $S_1$ を求めた。

##### [検討 3]

ベースとして【個人特性】と【追跡】のグループを考える。それらについて、今後の解析に支障のないと考えられる範囲で可能な限りセルの併合（まるめの処理）を行い、【個人特性】については2パターン、【追跡】については4パターンのサブグループを定義し、それらの分類数 $K$ とユニークセル数 $S_1$ を求めた。

#### [検討 4]

21 のアンケートカテゴリに対し アンケートカテゴリのみ、【個人特性】とアンケートカテゴリをそれぞれ組み合わせた場合、【追跡】とアンケートカテゴリをそれぞれ組み合わせた場合、【個人特性】【追跡】の組み合わせに各アンケートカテゴリを組み合わせた場合、の全ての場合における分類数 $K$ とユニークセル数 $S_1$ を求めた。

### C. 結果

[検討 1]より、単体の変数で一意である個体が存在するのは、「v0502 (10年観察終了日)」「v0600 (死因 ICD-9 コード 4 桁)」「v1200(身長 (cm))」「v1201 (体重 (kg))」「v1610 (初経年齢)」「v1612(自然閉経年齢)」「v1613(手術閉経年齢)」「v1615(出産人数)」「v1616(初産年齢)」「v2101 (喫煙開始年齢)」「v2102(喫煙本数/日)」「v2103 (禁煙年齢)」「v2801 (転入何年前か)」「v2940 (最も長く就いた仕事)」「v2950 (従事年数)」の15変数であった。(表2)

[検討 2]より、204 の全ての変数を組み合わせた場合に一意となる個体の数は100,605であった。

[検討 3]より、性別×年齢×居住地の情報からなる【個人特性】グループにおいて、まるめの処理を行わない「個人特性 1」では分類数 673、ユニークセル数 19 であったのに対し、年齢を5歳階級とし85歳以上はまとめた「個人特性 2」では、分類数 120、ユニークセル数は0と、一意性が消失

した(表3)。

追跡に関する日付×転帰×死因からなる【追跡】グループでは、処理を行わない「追跡 1」では分類数 20,176、ユニークセル数 16,631 であったのに対し、まるめの処理として、死因 ICD-9 コードを3桁までとする、かつ日付を月までにする(「追跡 2」)ことによりユニークセル数は約半分、同じく死因コード 3 桁かつ日付を追跡期間(単位:月)でみる(「追跡 3」)ことによりさらに半分になり、一意性は減少した。さらに死因情報を除いて日付を追跡期間(単位:月)で見た場合(「追跡 4」)では分類数が243、ユニークセル数が0になり一意性が消失した(表3)。

[検討 4] ~ の組み合わせから得られた329パターンについて、分類数、ユニークセル数、分類数に占めるユニークセル数の割合 $S_1 / K$ を示した(表4)。

また分類数 $K$ を横軸、ユニークセル数 $S_1$ を縦軸にその分布を示した(図1)。さらに、分類数 $K$ を横軸、分類数に占めるユニークセル数の割合 $S_1 / K$ を縦軸にその分布を示した(図2)。分類数が小さい時には分類数に占めるユニークセル数の割合も80%以下に分布するが、分類数の増加とともにユニークセルの割合が急増し、概ね分類数が20,000を超えると80%以上に分布した。すなわち、100,629例全体に対して16,000例程度(16%程度)が一意性のある個体数となり、分類数の増加に比例して、一意性のある個体数が増加した。

### D. 考察

各変数のユニークセル数の確認より、一意性には、変数 v0501 (10年観察終了日)のように、分類数が大きいことでそれぞれに振り分けられる個体数が少なくなるため生じる一意と、変数 v1615 (出産人数)において出産人数が20人というように、疫学的にまれな属性の個体が存在したために生じる一意の大きく2パターンが考えられた。前者に

対しては例えば日付データを月までにするなどにより分類数を減らすことで一意性を減少させることが可能であり、後者に対しては一定値以上（以下）については直接表示せず、無限までの片側区間で表示するといった方法により一意性の減少が図られる。

しかしながら今回すべての変数を組み合わせた場合の一意である個体の数は100,605であり、これは全レコード数の約99.98%にあたる。このように大規模なコホートデータにおいては、変数が多くなる（質問項目が多い）ことによる一意性は容易に避けられるものではない。また、本研究に利用した10万人規模のデータであるからまるめ処理などによりある程度の一意性の減少がみられるが、規模が小さくなると一意性が上がる可能性も高い。

分類数とユニークセル数の関係から、コホートデータにおいて、変数が増えるほど分類数は増大し、概ね分類数が20,000を超えると一意である個体の数も分類数の80%以上に分布した。一意性を上げないためには、一つのファイルに含む項目数を増やさないう、ファイルを分けて保管することなどが考えられるが、通常、一意性があるものとの前提で対応する必要がある。

死因に関しては、簡単分類を参考とした丸めの方法なども検討する必要がある。

## E. 結論

三府県コホートデータより、各変数、全変数あるいはいくつかの変数の組合せごとに一意性を検討した。三府県コホートデータのような10万人規模のデータの場合、分類数が概ね20,000を超えると一意性のある個体数は分類数の80%以上となり、一意性があるものとの前提で対応を考える必要がある。

## F. 研究発表

1. 論文発表
2. 学会発表

いずれもなし

## G. 知的財産権の出願・登録状況

（予定を含む。）

1. 特許取得
2. 実用新案登録
3. その他

いずれもなし

## Japan Arteriosclerosis Longitudinal Study (JALS)データにおける一意性の検討

研究分担者 大橋 靖雄 東京大学大学院医学系研究科  
研究協力者 原田亜紀子 東京大学大学院医学系研究科  
研究協力者 田島 里華 東京大学大学院医学系研究科

### 研究要旨

JALS のデータを用いて、死亡者の個人特定可能性について検討したところ、地域の情報が特に個人の特定に繋がりがやすいことが明らかとなった。さらに参加コホートが属する市町村の死亡集計表をもとにした約 30 万の死亡者を対象に検討結果では、少数セルの発生は 40～50 歳台で多くみられていたが、人口 30,000 人以上の市町村では総死亡においては、年齢階級（10 歳階級）別で少数セルの発生はみられなかった。集計対象を 60 歳以上に限定した場合では、5,000 人以上の市町村であれば総死亡、循環器疾患死亡ともに少数セルの発生を抑えられる可能性が考えられた。

### A. 目的

Japan Arteriosclerosis Longitudinal Study (JALS) 対象者で死亡したものに付き、特定の変数による個人特定可能性を検討する。さらに、JALS 参加コホートが属する市町村の総死亡、循環器疾患死亡について、男女別に 10 歳階級別で死亡者数を集計した場合に、どのような条件において少数例（セル）が発生するかを検討する。

### B. 方法

#### ・ JALS 対象者における死亡者に関する検討

JALS 対象者(年齢は 40～89 歳に限定)のうち、各コホートが住基情報等をもとに 2010 年度までに特定した死亡者 4,858 人について、特定の変数による個人特定可能性を検証した。

#### ・ JALS 対象地域の死亡数と市町村規模の検討

JALS 参加コホートが属する市町村について、年齢階級別の死亡集計表をもとに、1999 年から 2010 年までの 12 年間で市町村合併のなかった 88 市町村 40 歳以上の死亡者 295,846 例を分析対象に、市町村規模と特定可能性の検討を行った。40 歳以上の集団で検討した場合と死亡数が多い 60 歳以上に限定した場合を検討した。

### C. 結果

#### ・ JALS 対象者における死亡者に関する検討

##### 1. 単変数による個人特定可能性の検討

##### 1-1. 性別

男性 2,967 人 (61.1%)、女性 1,891 人 (38.9%) であり、性別のみからは個人は特定されない。

##### 1-2. ベースライン時年齢

図 1 より、ベースライン時年齢ごとの人数は最少でも 6 人であり、この変数のみにより個人を特定されない。

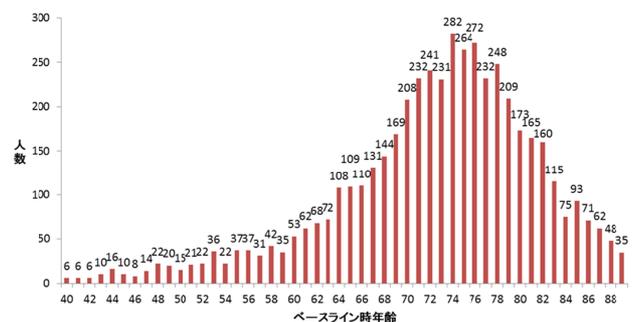


図 1. ベースライン時年齢ごとの人数分布

##### 1-3. 死亡時年齢

図 2 と表 1 より、死亡時年齢が 40 代前半もしくは 90 代後半の場合、死亡時年齢のみから個人を特定できる可能性は高くなる。

表1. 死亡者数が5人以下であった死亡時年齢

死亡時年齢	人数
44	3
45	2
96	5
98	2
99	1

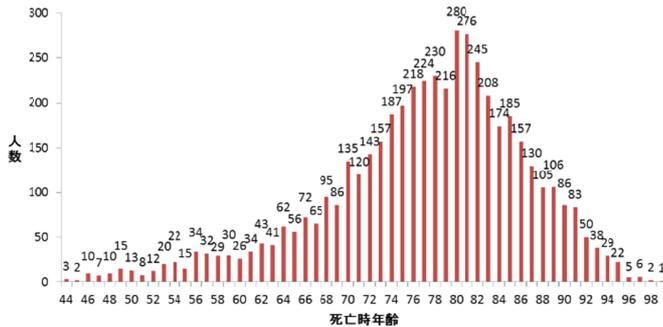


図2. 死亡時年齢の分布

1-4. 地域 (コホート・市町村)

コホート単位で死亡者を集計したところ、死亡者数が20名以下のコホートが2つ(12人と17人)存在した。またコホート単位に比べより詳細な地域コード(市町村)ごとにも検討したところ、死亡者数が1名のみ、10名以下、20名以下の地域がそれぞれ3か所、9か所、18か所存在していた(表2)。地域を公表する際にはコホート単位にとどめることにより、地域情報のみからの個人特定可能性を下げる可以考虑とされる。

地域コード	人数
33-3	1
13-11	1
33-14	1
33-6	2
33-2	3
33-1	4
33-13	7
33-4	8
2-5	9
3-17	11
21-1	12
3-13	12
3-16	13
29-1	17
2-1	17
3-11	18
1-2	19
3-6	19

表2. 死亡者数が20人以下となる地域

1-5. 死亡年月日

死亡年月日ごとの人数を集計した結果を表3に示した。全死亡者4,858人のうち、21.2%にあたる1,028人はユニークな死亡年月日を有していた(表3)。また、同一死亡年月日を持つ人が5人以下である場合がほとんどであるため、この変数は特定性が高いと考えられた。死亡情報を死亡年月までに丸めて集計した結果を表4に示した。死亡年月までの丸めを行っても、個人が一意に特定されるパターンが10個存在し、これ以外にも同一死亡年月を持つ人数が10人以下のパターンも複数存在していた。よって死亡時点を公表する際には、死亡年まで丸めて公表するなど、データを粗くする必要が考えられた。

表3. 死亡年月日の特異性の検討

同一死亡日を持つ人数(人)	死亡日のパターン数	合計人数(人)	割合(%)
1	1028	1028	21.2
2	648	1296	26.7
3	361	1083	22.3
4	160	640	13.2
5	69	345	7.1
6	27	162	3.3
7	15	105	2.2
8	8	64	1.3
9	5	45	0.9
10	2	20	0.4
11	4	44	0.9
12	1	12	0.2
14	1	14	0.3

表4. 死亡年月の特異性の検討

同一死亡年月を持つ人数(人)	死亡年月のパターン数
1	10
2	7
3	3
4	5
5	5
6	1
7	1
8	2
9	1

\*地域コード：コホート番号 - コホート内の市町村通番

## 2.2 変数による個人特定可能性の検討

### 2-1. 性とベースライン時年齢

表 5 より、40 代の対象者においては、性別との組み合わせにより個人の特定可能性が高いが、ベースライン時年齢を 5 歳刻みのカテゴリ変数にすることなどで対処可能であると考えられる。

**表 5. 人数が 5 人以下となる性・ベースライン年齢の組み合わせ**

性	ベースライン時年齢	人数
男性	40	2
	41	2
	42	4
	47	4
女性	40	4
	41	4
	42	2
	43	2
	45	4
	46	1
	51	4

### 2-2. 性と死亡時年齢

表 6 より、男性では死亡時年齢が 40 代と 90 代、女性では 40, 50 代と 90 代では、性別と死亡時年齢を組み合わせること個人が特定される可能性が高くなり、一意に特定できるパターンも 2 つ存在する。

**表 6. 人数が 5 人以下となる性・死亡時年齢の組み合わせ**

性別	死亡時年齢	人数
男性	44	2
	45	2
	47	4
	48	5
	96	3
	97	4
女性	44	1
	46	4
	47	3
	48	5
	50	4
	51	2
	52	5
	55	5
	96	2
	97	2
	98	2
	99	1

**表 7. 個人が一意に特定できるベースライン時と**

### 2-3. 性と地域

地域情報の公開をコホートレベルにとどめ、性別と組み合わせた場合、合計人数が 5 以下となるパターンは、女性に関して、2 人のみという地域が 1 か所存在した。

### 2-4. ベースライン時年齢と死亡時年齢

ベースライン時と死亡時の年齢を組み合わせた情報は個人特定に繋がりがやすく、78 人において一意に個人を特定することが可能であった（表 7）。どちらかの年齢が比較的若いもしくは高齢である場合、または死亡までの年数が短い場合に特定されやすいことも明らかになった。両変数ともに 5 歳刻みにカテゴリ化することで、死亡時年齢が欠損している 1 例を除き、個人の特定可能性について対象可能であった（表 8）。

### 2-5. ベースライン時年齢と地域

ベースライン時年齢（1 歳刻み）とコホートを組み合わせると、200 人の個人を一意に特定することが可能であった。年齢を 5 歳刻みにしたところ、この人数は 27 人となった。同様に年齢を 10 歳刻みとした場合に、人数が 5 人以下となるパターンを表 9 に示した。ベースライン時年齢が比較的若い場合、またコホート内の死亡が少ない場合に、これら 2 つの情報を組み合わせること個人が特定可能性が上昇すると考えられた。

### 2-6. ベースライン時年齢と死亡年

1 歳刻みのベースライン時年齢とコホートコードを組み合わせると、67 人が一意に特定された。年齢を 5 歳刻みにした場合、人数が 5 人以下となる組み合わせを表 10 に示した。ベースライン時年齢と死亡年を組み合わせた場合、年齢カテゴリを荒くすることのみでは個人特定を防ぐことができず、特に追跡早期に死亡した例で特定可能性が高かった。

**個人が一意に特定できるベ**

**ースライン時と**

### 死亡時年齢の組み合わせ

Index	ベースライン時年齢	死亡時年齢	Index	ベースライン時年齢	死亡時年齢
1	40	46	40	59	66
2	40	47	41	59	68
3	41	44	42	60	69
4	41	48	43	61	61
5	42	46	44	61	70
6	42	48	45	62	71
7	43	51	46	64	64
8	44	46	47	65	65
9	44	53	48	66	66
10	45	47	49	68	68
11	45	48	50	72	82
12	45	49	51	73	73
13	45	50	52	73	82
14	45	52	53	74	83
15	45	53	54	74	84
16	46	47	55	75	86
17	46	50	56	77	77
18	48	51	57	77	87
19	49	50	58	77	88
20	49	51	59	78	88
21	49	55	60	78	90
22	49	57	61	79	89
23	49	58	62	79	90
24	50	56	63	80	90
25	50	58	64	80	91
26	50	59	65	83	92
27	51	52	66	85	85
28	51	55	67	85	95
29	51	59	68	85	96
30	52	53	69	85	97
31	52	60	70	86	95
32	54	55	71	86	96
33	54	63	72	86	97
34	55	55	73	87	87
35	55	64	74	87	97
36	57	57	75	88	96
37	57	66	76	88	98
38	58	58	77	89	98
39	59	60	78	89	99

### み合わせ

ベースライン時年齢カテゴリ	コホートコード	人数
40	10	1
40	23	1
40	31	1
ベースライン時年齢カテゴリ	死亡年	人数
40	2003	1
40	2005	1
45	2000	1
55	2000	1
60	2000	1
70	.	1
80	1999	1
40	2010	2
65	2000	2
70	1999	2
70	2000	2
85	2001	2
40	2004	3
60	2001	3
65	2001	3
50	2002	4
40	2006	5
45	2003	5
50	2003	5
55	2002	5
80	2000	5
70	21	5

表 10.人数が 5 人以下となるベースライン時年齢（10 歳刻み）と死亡年の組み合わせ

表 8.人数が 5 人以下となるベースライン時と死亡時年齢（5 歳刻み）の組み合わせ

ベースライン時年齢カテゴリ	死亡時年齢	人数
40	40	3
70	.	1
75	90	2

表 9.人数が 5 人以下となるベースライン時年齢（10 歳刻み）と地域の組

## 2-7. 死亡時年齢と地域

死亡時年齢（1歳刻み）とコホートを組み合わせると、191人の個人が一意に特定された。年齢を5歳刻みにしたところ、この人数は32人となった。年齢を10歳刻みとした場合に、人数が5人以下となるパターンを表11に示した。死亡が少ないコホートにおいて比較的若いまたは高齢の人が死亡した場合で、特定可能性が上昇していた。

**表 11. 人数が5人以下となる死亡時年齢（10歳刻み）と地域の組み合わせ**

死亡時年齢 カテゴリ	コホート コード	人数
40	4	1
40	8	1
40	10	1
40	20	1
40	23	1
40	31	1
40	34	1
50	1	1
50	20	1
50	21	1
50	25	1
50	29.1	1
50	34	1
60	29.1	1
90	21	1
90	31	1
90	33	1
40	13	2
40	17	2
40	30	2
60	21	2
60	27	2
90	29.1	2
40	2	3
40	5	3
50	33	3
80	21	3
90	23	3
40	7	4
50	4	4
50	8	4
50	23	4
50	31	4
50	2	5
60	1	5
70	21	5
90	8	5
90	11	5
90	34	5

## 2-8. 死亡時年齢と死亡年

死亡時年齢（1歳刻み）とコホートを組み合わせることで、74人が一意に特定された。年齢を5歳刻みにした場合で、人数が5人以下となる組み合わせを表12に示した。特に追跡早期の死亡例で特定性が高く、死亡時年齢だけでなく、死亡年を丸めるなどの対処が必要であると考えられた。

**表 12. 人数が5人以下となる死亡時年齢（5歳刻み）と死亡年の組み合わせ**

死亡時年齢 カテゴリ	死亡年	人数
40	2003	1
40	2004	1
40	2007	1
45	2000	1
55	2000	1
60	2001	1
70	2000	1
80	1999	1
90	2001	1
95	2004	1
45	2010	2
50	2010	2
65	2001	2
70	1999	2
85	2000	2
95	2006	2
45	2005	3
50	2002	3
65	2000	3
80	2000	3
85	2001	3
95	2007	3
45	2004	4
50	2003	4
60	2002	4
45	2003	5
55	2002	5
95	2008	5

### 3. 3 変数による個人特定可能性の検討

#### 3-1. 性、ベースライン時年齢、死亡時年齢

これら3つの変数を組み合わせる際には、局所的にカテゴリを粗くする等、個別例に関する検討が必要であると考えられた(表13)。

**表13. 人数が5人以下となる性、ベースライン時年齢(5歳刻み)、死亡時年齢の組み合わせ**

性別	ベースライン時 年齢カテゴリ	死亡時年齢 カテゴリ	人数
男性	40	40	2
	70	.	1
	75	90	1
女性	40	40	1
	40	50	3
	45	45	5
	45	55	5
	50	50	5
	50	60	5
	75	90	1
性別	死亡時年齢 カテゴリ	死亡年	人数
男性	40	2000	1
	40	2005	1
	50	2000	1
	80	2000	1
	90	2001	1
女性	60	2001	1
	80	1999	1
	80	2001	1
男性	60	2001	2
	70	1999	2
女性	40	2005	2
	40	2006	2
	40	2010	2
男性	90	2003	2
	40	2003	3
	40	2008	3
女性	60	2000	3
	40	2003	3
	40	2008	3
	40	2009	3
	50	2010	3
男性	70	2000	3
	40	2007	4
	40	2009	4
	50	2002	4
女性	90	2003	4
	50	2002	4
	60	2002	4
	80	2000	4
男性	40	2004	5
	70	2000	5
	90	2004	5
女性	40	2007	5
	50	2003	5

#### 3-2. 性、死亡時年齢、地域

これら3変数の組み合わせ(死亡時年齢は5歳刻み)より、76人について一意の特定が可能であった。死亡時年齢を10歳刻みにしても一意に特定可能であった32例について表14に示した。死亡が少ないコホートについては変数の情報が加わるほど個人が特定される可能性が大きくなるため、情報の開示の可否も含め検討する必要がある。開示を行う場合には近接地域で統合し開示する、地域を数字や記号などを用い秘匿するなどといった対策が必要になると考えられる。

#### 3-3. 性、死亡時年齢、死亡年

死亡時年齢を5歳刻みとしたとき、これら変数の組み合わせより一意に特定できるのは23名であった。10歳刻みとした場合の結果を表15に示した。死亡年を複数年でまとめることで個人特定の可能性を低くすることが可能と考えられた。

**表14. 個人が一意に特定される性、死亡時年齢(10歳刻み)、地域の組み合わせ**

#### 3-4. 性、地域、死亡年

3-2同様に、コホートは他に2つの変数と組み合わせると個人特定能が高く、この3変数の組み合わせからは43例が一意に特定可能であった。開示にあたっては、死亡年をさらに粗くカテゴリ化するか、もしくは地域に関して3-2で示した措置を施す必要がある。

**表15. 人数が5人以下となる性、死亡時年齢(10歳刻み)、死亡年の組み合わせ**

#### 3-5. ベースライン時年齢、死亡時年齢、地域

年齢についての2変数をそれぞれ5歳刻みにすると、95人を一意に特定することが可能であった。

同様に 10 歳刻みにし、一意に個人が特定できる例を表 16 に示した。特定可能例を多く含むコホートがいくつか存在しており、この結果からも 3-2 に示した地域情報に対する措置が必要であると考えられた。

**表 16 一意に特定可能なベースライン時と死亡時年齢  
(10 歳刻み)、地域の組み合わせ**

3-6. ベースライン時年齢、死亡時年齢、死亡年  
ベースライン時年齢と死亡時年齢を 5 歳刻みにすることにより一意特定できる個人は 27 人であった。同様に両変数を 10 歳刻みにした場合に、合計人数が 5 人以下であった変数のパターンを表 17 に示した。ベースライン時および死亡時年齢のカテゴリ化に加え、死亡年を 5 年刻みでカテゴリ化することで特定可能性を小さくすることが可能と思われた。

3-7. ベースライン時年齢、地域、死亡年  
この変数の組み合わせでは、ベースライン時年齢を 10 歳刻みにした場合でも、149 人が一意に特定可能であった。表 18 に死亡年を 5 年刻みにしても一意に特定可能であった 39 例を示した。コホートの情報は他変数と組み合わせる際に注意が必要であることが考えられる。

ベースライン時 年齢カテゴリ	死亡時年齢 カテゴリ	コホートコード
40	40	4
40	40	8
40	40	10
40	40	20
40	40	23
40	40	31
40	40	34
40	50	5
40	50	8
40	50	17
40	50	20
40	50	33
50	50	1
50	50	21
50	50	25
50	50	29.1
50	50	34
50	60	23
60	60	29.1
60	70	21
70	.	33
70	80	21
80	90	21
80	90	31
80	90	33

**表 17. 人数が5人以下となるベースライン時と死亡時年齢（10歳刻み）、死亡年の組み合わせ**

ベースライン時 年齢カテゴリ	死亡時年齢 カテゴリ	死亡年	人数
40	40	2000	1
50	50	2000	1
50	60	2002	1
70	.	.	1
70	90	2009	1
70	90	2010	1
80	80	1999	1
80	90	2001	1
40	40	2010	2
50	60	2003	2
70	70	1999	2
40	40	2005	3
60	60	2000	3
60	60	2001	3
60	70	2001	3
60	70	2002	3
40	50	2004	4
40	50	2005	4
50	50	2010	4
70	80	2001	4
40	40	2004	5
40	50	2006	5
80	80	2000	5

3-8. 死亡年齢、地域、死亡年

コホート変数は、3変数以上の組み合わせで個人の特定期間可能性が高くなり、死亡年を複数まとめてカテゴリ化しても個人特定性は高かった。表 19 に死亡時年齢を 10 歳刻み、死亡年を 5 年刻みとして、一意特定が可能であった例を示した。

**表 18. 一意特定可能なベースライン時年齢（10歳刻み）と地域、死亡年の組み合わせ**

ベースライン時 年齢カテゴリ	コホート コード	死亡年
40	2	2005
40	3	2000
40	4	2000
40	5	2000
40	8	2005
40	8	2010
40	10	2005
40	13	2000
40	20	2000
40	20	2010
40	23	2005
40	28	2000
40	28	2010
40	30	2010
40	31	2005
40	33	2005
40	34	2010
50	1	2000
50	4	2000
50	20	2005
50	21	2005
50	23	2010
50	25	2000
50	29.1	2000
50	33	2010
60	4	2010
60	6	2000
60	21	2000
60	22	2010
70	8	2000
70	20	2000
70	21	2000
70	33	.
70	33	2010
80	23	2010
80	27	1995
80	29.1	2000
80	31	2010
80	33	2000

表 19. 一意特定可能な死亡時年齢（10 歳刻み）と地域、死亡年の組み合わせ

死亡時年齢	コホート コード	死亡年	死亡時年齢	コホート コード	死亡年
40	2	2005	60	8	2010
40	3	2000	60	10	2010
40	4	2000	60	21	2000
40	5	2000	60	21	2005
40	7	2005	60	25	2010
40	8	2005	60	29.1	2005
40	10	2005	60	33	2000
40	13	2005	70	4	2010
40	13	2010	70	8	2000
40	20	2000	70	21	2000
40	23	2005	80	20	2000
40	31	2005	80	27	1995
40	34	2010	80	33	2000
50	1	2000	80	33	2010
50	4	2000	90	2	2000
50	8	2010	90	5	2000
50	17	2000	90	6	2000
50	17	2010	90	10	2000
50	20	2010	90	17	2000
50	21	2005	90	21	2005
50	25	2000	90	23	2010
50	29.1	2000	90	31	2010
50	30	2010	90	33	2005
50	33	2010	90	34	2010
50	34	2005			

組み合わせを検討した場合に、人数が 2 人以下

#### 4.4 変数による個人特定可能性の検討

前節での 3 変数の検討結果より、コホート変数を他の変数と組み合わせると特定可能性が高くなることが明らかになった。さらに変数を 1 つ加える検討として、4-1 では地域情報を含まない組み合わせにおいて、各変数においてカテゴリ化をどの程度粗くすることで個人の特定を防ぐことができるかを検討する。4-2~4-5 では、コホート変数に 3-2 のような処理を行わない場合に、他の変数のカテゴリ化をどの程度粗くすることで個人の特定可能性が変化するかを検討する。

##### 4-1. 性、ベースライン時年齢、死亡時年齢、死亡年

2 つの年齢の変数を 10 歳刻みとし、4 つの変数の

となる組み合わせパターンを表 20 に示した。さらに、死亡年を 5 年刻みのカテゴリ化の処理を行い組み合わせた場合を表 21 に示した。この条件では、人数が 5 人以下となるパターンは 11 個、うち一意に特定される人は 5 パターンであった。

##### 4-2. 地域、性、ベースライン時年齢、死亡時年齢

両年齢変数を 5 歳刻み、10 歳刻みにした場合、一意に特定できる人数は 226 人、58 人であった。

**表 20. 人数が 2 人以下となる性、ベースライン時、死亡時年齢（10 歳刻み）、死亡年の組み合わせ**

性別	ベースライン時 年齢カテゴリ	死亡時年齢 カテゴリ	死亡年	人数
男性	40	40	2000	1
	40	40	2005	1
	40	50	2006	1
	50	50	2000	1
	50	60	2002	1
	50	60	2003	1
	60	70	2002	1
	70	.	.	1
	70	90	2010	1
	80	80	2000	1
女性	80	90	2001	1
	40	50	2004	1
	40	50	2005	1
	50	60	2003	1
	60	60	2001	1
	70	90	2009	1
男性	80	80	1999	1
	80	80	2001	1
	60	60	2001	2
女性	70	70	1999	2
	70	80	2002	2
	40	40	2005	2
	40	40	2006	2
	40	40	2010	2
	50	50	2009	2
	50	60	2005	2
60	70	2002	2	
80	90	2003	2	

**表 21. 人数が 5 人以下の性、ベースライン時、死亡時年齢（10 歳刻み）、死亡年（5 歳刻み）の組み合わせ**

性別	ベースライン時 年齢カテゴリ	死亡時年齢 カテゴリ	死亡年	人数
男性	70			1
男性	70	90	2010	1
女性	40	50	2000	1
女性	70	90	2005	1
女性	80	80	1995	1
男性	70	70	1995	2
女性	40	40	2010	2
男性	40	50	2000	3
女性	40	40	2000	3
女性	40	50	2010	3
男性	50	50	2010	4

4-3. 地域、性、ベースライン時年齢、死亡年  
死亡年が 1 年刻みの場合、ベースライン時年齢が 5 歳、10 歳刻みとした場合に一意に特定される人数は 638 人、318 人であった。また死亡年を 5 年刻みのカテゴリ変数とした場合、年齢が 5 歳、10 歳刻みのときに一意に特定される人数は 176 人、90 人であった。

4-4. 地域、性、死亡時年齢、死亡年  
死亡年が 1 年刻み、死亡時年齢が 5 歳、10 歳刻みの場合に一意に特定される人数は 682 人、366 人であった。また死亡年を 5 年刻みのカテゴリ変数、年齢が 5 歳、10 歳刻みのときに一意に特定される人数は 214 人、110 人であった。

4-5. 地域、ベースライン時年齢、死亡時年齢、死亡年  
死亡年が 1 年刻みの場合、2 つの年齢変数を 5 歳、10 歳刻みとした場合に一意に特定される人数は 614 人、309 人であった。また死亡年を 5 年刻みのカテゴリ変数とした場合、両年齢変数が 5 歳、10 歳刻みのときに一意に特定される人数は 228 人、90 人であった。

5. 5 変数による個人特定可能性の検討  
年齢変数を 10 歳刻み、死亡年を 5 年刻みのカテゴリ変数とし、今回検討した全 5 変数を組み合わせた際に、一意に特定された人は 186 人であった（表 22, 23）。これには、地域の情報が大きく影響しており、地域情報を公開する際には工夫が必要である。

表 22 一意に特定される 5 変数の組み合わせ (男性)

コホート コード	ベースライン時 年齢カテゴリ	死亡時年齢 カテゴリ	死亡年	コホート コード	ベースライン時 年齢カテゴリ	死亡時年齢 カテゴリ	死亡年
1	60	70	2000	21	60	60	2000
1	60	70	2005	21	60	60	2005
2	40	40	2005	21	60	70	2005
2	50	60	2005	21	70	70	2000
2	50	60	2010	21	70	80	2005
2	70	70	2010	21	80	90	2005
2	80	80	2010	22	60	70	2000
2	80	90	2000	22	70	70	2010
3	40	40	2000	22	80	80	2000
4	40	40	2000	22	80	80	2010
4	40	50	2005	22	80	90	2010
4	50	50	2000	23	80	90	2005
4	50	50	2005	25	50	50	2000
4	60	70	2010	25	60	60	2010
4	70	80	2000	25	60	70	2010
5	40	40	2000	25	70	70	2010
5	40	50	2005	25	70	80	2000
5	50	60	2000	27	60	60	2005
5	60	70	2000	27	60	70	2010
5	70	80	2000	27	70	90	2010
6	60	70	2000	28	40	50	2000
6	80	90	2000	28	50	50	2010
7	40	40	2005	28	50	60	2000
8	40	40	2005	28	60	70	2000
8	50	50	2005	29.1	50	50	2000
8	60	70	2010	29.1	70	80	2005
10	60	70	2000	29.1	80	80	2000
10	60	70	2010	29.1	80	80	2005
10	80	80	2000	30	40	50	2010
13	40	50	2000	30	50	50	2000
13	50	50	2000	30	50	50	2005
13	50	60	2000	30	50	60	2000
13	50	60	2005	30	50	60	2010
13	70	80	2000	31	40	40	2005
14	40	50	2000	31	50	60	2005
14	50	50	2010	31	80	90	2010
17	50	50	2010	33	50	50	2005
17	50	60	2000	33	50	50	2010
17	70	70	2010	33	50	60	2005
17	80	90	2000	33	60	70	2000
20	40	40	2000	33	70		
20	50	60	2005	33	70	80	2010
20	50	60	2010	33	80	80	2000
20	60	60	2000	33	80	90	2005
20	70	80	2000	34	50	50	2005
20	80	80	2010	34	60	70	2010
20	80	90	2005	34	80	90	2010

表 23 一意に特定される 5 変数の組み合わせ (女性)

コホート コード	ベースライン時 年齢カテゴリ	死亡時年齢 カテゴリ	死亡年	コホート コード	ベースライン時 年齢カテゴリ	死亡時年齢 カテゴリ	死亡年
1	50	50	2000	20	40	50	2010
1	60	60	2000	20	50	60	2010
1	60	70	2000	20	60	60	2000
1	80	80	2005	20	60	60	2005
2	50	60	2000	20	60	70	2010
2	50	60	2005	21	50	50	2005
2	50	60	2010	21	70	70	2005
2	80	80	2000	22	50	50	2005
2	80	90	2005	22	60	70	2010
3	50	50	2000	22	80	90	2010
4	40	50	2005	23	40	40	2005
4	60	60	2000	23	50	50	2005
4	70	70	2000	23	50	60	2010
4	70	80	2000	23	60	60	2010
4	70	80	2010	23	60	70	2010
5	50	50	2000	23	70	70	2000
5	70	70	2000	23	80	90	2005
5	80	90	2000	23	80	90	2010
6	60	70	2005	25	50	60	2005
7	40	40	2000	25	60	70	2000
8	40	50	2010	27	60	60	2005
8	60	60	2005	27	70	70	2010
8	60	60	2010	27	70	90	2005
8	60	70	2005	27	80	80	1995
8	60	70	2010	28	40	50	2010
8	70	70	2000	28	50	50	2005
8	70	70	2010	28	60	60	2010
8	70	80	2010	28	80	90	2000
10	40	40	2005	29.1	60	60	2005
10	60	60	2010	29.1	70	70	2005
10	70	70	2010	29.1	70	80	2000
10	80	90	2000	30	50	50	2000
13	40	40	2005	30	50	60	2000
13	40	40	2010	31	50	60	2005
13	50	60	2010	31	60	60	2010
13	60	60	2000	33	40	50	2005
13	70	70	2010	33	50	60	2005
13	70	80	2000	33	60	60	2000
13	80	80	2010	34	40	40	2010
14	40	50	2000	34	50	60	2005
17	40	50	2005	34	60	70	2010
17	50	50	2000	34	70	70	2005
17	50	60	2000	34	80	90	2005
17	60	60	2000				
17	60	70	2000				
17	70	70	2000				
17	70	70	2010				
17	80	80	2010				
17	80	90	2010				

## ・ JALS 対象地域の死亡数と市町村規模の検討

### 1 . 少数セルの発生状況

40 歳以上での検討では、 総死亡では、 18 市町村、 CVD 死亡では、 29 市町村において、 10 未満の少数セルが発生した。

上記の検討を死亡数の多い 60 歳以上の対象に限定して同様に行うと、 総死亡では、 2 市町村、 CVD 死亡では、 9 市町村において少数セルが発生した。 40 歳以上の集団での検討に比べて、 10 未満の少数セルの発生は、 総死亡、 CVD 共に減少した。

### 2 . 市町村規模での検討

#### 1 ) 40 歳以上での検討

市町村について、 人口規模 ( -4,999, 5,000-9,999, 10,000-29,999, 30,000-49,999, 50,000-)に分け、 少数セルの発生パターンを検討した ( 表 23 )。

a:-4999 ( 9 市町村 ) b:5000-9999 ( 5 市町村 ) では、 総死亡、 CVD 死亡ともに集計区分で少数セルが発生していた。 c:10000-29999 ( 8 市町村 ) では、 CVD の区分で少数セルが発生していたが、 総死亡の区分では少数セルが発生したのは 4 市町村であった。 d:30,000-49,999 ( 5 市町村 ) e:50,000-( 4 市町村 ) では、 CVD 死亡では大部分の市町村で少数セルが発生していたが、 総死亡では発生がみられなかった。

#### 2 ) 60 歳以上での検討

1 ) と同様の検討を死亡数の多い 60 歳以上の対象に限定して行うと、 5000 人以上の市町村であれば、 総死亡、 CVD 死亡ともに少数セルの発生を抑えられる可能性が考えられた。

## E . 結論

JALS 対象者のうち死亡例について、 データ開示による個人の特定の可能性を検討した。 検討した変数のうち、 地域の情報が特に個人の特定の繋がりやすいことが明らかとなった。 地域に関する情報を公表する場合には、 事前に何らかの工夫を施す必要があると考えられた。

また、 約 30 万の死亡者を対象に市町村規模を考慮して検討した結果では、 少数セルの発生は、 40 ~ 50 歳台で多くみられていたが、 人口が 30,000 人規模以上であれば、 総死亡においては少数セルの発生はみられなかった。 集計対象を死亡数も多くなる 60 歳以上に限定することにより、 人口規模が 5,000 人以下の村レベルを除けば、 人口規模によらず総死亡、 CVD とともに少数セルの発生数を抑えられる可能性が考えられた。

## F . 研究発表

1. 論文発表
  2. 学会発表
- いずれもなし

## G . 知的財産権の出願・登録状況

( 予定を含む。 )

1. 特許取得
  2. 実用新案登録
  3. その他
- いずれもなし

表 23 市町村規模と死亡数の関連

コホート 番号	人口規模			40 歳以上での検討		60 歳以上での検討	
	人口カテゴリー	人口	75 歳以上 人口割合	総死亡	CVD 死亡	総死亡	CVD 死亡
				10 歳階級	10 歳階級	10 歳階級	10 歳階級
27		1,380	23.0				
25		2,573	20.4				
1		2,995	16.8				
8		3,150	16.0				
28	a: -4,999	3,920	15.2				
7		4,072	17.8				
24		4,629	20.0				
29		4,683	22.2				
10		4,936	15.2				
26		5,243	16.9				
13		5,706	15.9				
11	b: 5,000-9,999	6,789	18.1				
14		7,144	13.4				
20		8,283	11.7				
9		10,868	16.4				
23		11,132	18.5				
6		11,489	19.1				
22	c: 10,000-29,999	11,674	13.7				
19		11,920	22.5				
12		14,931	18.6				
30		17,366	16.6				
5		19,584	14.8				
4		30,988	15.4				
15		36,379	17.0				
3	d: 30,000-49,999	38,569	12.7				
18		39,981	16.8				
31		47,973	8.5				
21		52,156	14.9				
2	e: 50,000-	57,912	14.2				
16		65,163	21.2				
17		125,515	8.3				

: 観察数が 10 未満のセルが発生

## 疫学研究データのアーカイブ化の試み

研究分担者	辻一郎	東北大学大学院医学系研究科
研究分担者	祖父江友孝	大阪大学大学院医学系研究科
研究代表者	玉腰暁子	北海道大学大学院医学研究科
研究協力者	山縣 然太郎	山梨大学大学院医学工学総合研究部

### 研究要旨

日本疫学会統計利用促進委員会と共同で、疫学研究データをアーカイブ化し、外部に委託して管理・運営する際に定めておくべき内容を、特に共同研究の場合を念頭に、管理者、データ提供者、データ利用者別に整理した。疫学データの個別性、データの持つ背景要因の理解の困難さを考えると、データを全ての人に対しオープンにすることが必ずしも好ましい結果をもたらさない可能性も考えられる。一方で、多くの疫学研究のデータは、多額の費用と多くの人々の協力、長期にわたる研究により得られた貴重な情報である。適切なデータ提供ならびに利用のあり方について、今後も慎重な議論が必要である。

### A. 目的

疫学研究で収集された個人単位のデータについて、そのアーカイブ化と利用体制の構築を目指し、試験的な実施を試みる。

### B. 方法

日本疫学会統計利用促進委員会(委員長:山縣然太郎)と共同し、外部に委託してデータアーカイブを管理・運営する際に定めておくべき内容を整理する。

### C. 結果

既存の疫学研究データをアーカイブ化し、一定のルールの下で公開する目的は、大きく、

1. 公費等を投入し、多くの人々の協力を得て作られた貴重な疫学データの有効活用
2. 若手研究者の育成

### 3. データの検証

に分けることができる。また、公開の範囲も、無条件に全ての項目を全ての人にオープンにするレベルから、一定の審査等手続きを経て承認された研究者に対して一定の項目を提供するレベルまであり、後者も条件をつけない提供レベルから、共同研究まで、多くの段階が考えられる。そこで、今回は、まずデータ情報を公開することにより、共同研究を行える体制を構築するための要件を検討した。

[管理者]

制度を適切に運営するために、データの管理責任者を置く。さらに、提供されるデータの受け入れ、利用申請の整理と承認等を行うための運営委員会・事務局が必要と考えられる。提供されたデータについては、利用希望者が利用を検討できる範囲での情報をネット等で公開する。提供されるデータ数の増加、利用希

望者の増加に伴い、データ預かり業務や事務作業が増えることが予想される。金銭的な手当についても今後検討が必要と考えられる(例えば、利用者から一定金額を徴収するなど)。

[データ提供者]

データ提供者には、法的、倫理的に問題のないデータを提供できるよう、インフォームドコンセントの範囲の確認、ならびに所属機関での倫理審査を求める。また、共同研究利用の申し込みを検討するために必要な情報公開の内容としては、少なくとも、研究の概要、調査方法、母集団、標本数、有効回答率、データ収集時期、項目数、ファイル形式を挙げることができた。また、データの詳細を示す情報(調査票、項目名と変数等)の提供も必要である。カテゴリ共同研究のあり方として、データ提供者の承諾が必要か不要か、結果公表時のオーサiershipに対する希望等に関しても、あらかじめ意思表示を求めることが望ましい。

[データ利用者]

今回は二次利用ではあるが共同研究の位置づけを想定している。利用者は公開されている調査内容に基づき、共同研究の申し入れを行う。データ利用にあたっては、提供者との共同研究利用にかかる契約を交わすとともに、ルールに基づいた誓約書を提出する。なお、利用者は研究成果を論文等で公表しなくてはならない。

これらに関し、運用のためのマニュアルを整備することが必要と考えられた。

#### D. 考察

疫学研究で得られたデータをアーカイブ化し二次利用体制を整備するため、まず共同研究の場合を例に、検討すべき内容を整理した。

公的研究費により実施されるライフサイエンス分野の研究では、現在、論文発表等で公表された成果に関わる生データの複製物、又は構築した公開用データベースの複製物を、バイオサイエンスデータベースセンターに提供することが求められている。しかし、疫学データの個別性、データの持つ背景要因の理解の

困難さを考えると、データを全ての人に対しオープンにすることが必ずしも好ましい結果をもたらさない可能性も考えられる。多くの疫学研究のデータは、多額の費用と多くの人々の協力、長期にわたる研究により得られた貴重な情報である。適切なデータ提供ならびに利用のあり方について、今後も慎重な議論が望まれる。

#### E. 結論

日本疫学会統計利用促進委員会と共同で、疫学研究データをアーカイブ化し、外部に委託して管理・運営する際に定めておくべき内容を、特に共同研究の場合を念頭に整理した。

#### F. 研究発表

1. 論文発表
  2. 学会発表
- いずれもなし

#### G. 知的財産権の出願・登録状況

(予定を含む。)

1. 特許取得
  2. 実用新案登録
  3. その他
- いずれもなし

研究成果の刊行に関する一覧表

2013 年度：なし