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25～27年度

総合研究報告書

平成 28（2016）年3月

研究代表者　玉腰　暁子

# 目　次

厚生労働省科学研究費補助金(循環器疾患・糖尿病等生活習慣病対策総合研究事業)
総括研究報告書

# 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

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**

疫学研究により得られたデータを広く共有化するためのシステムであるデータアーカイブ化に向けた検討を行った。統計法の規定上、死因情報を付加してのデータ公開・二次利用は認められていないことなどを踏まえて、現制度下での運用方法を2つ提案した。また、実データを用いて、一意性に関する検討を行い、10万人規模のコホート集団はもちろん、100例規模のコホート集団でも分類数が全対象者数の40%程度で、ユニークセルの割合が80%に達していたことから、コホートの規模にかかわらず、80%程度のレコードは一意性があるものとして対応する必要があることを示した。これら研究成果や他国・他領域での実例を参考に、データアーカイブの利活用を進めるため、データ提供の際に従うべきガイダンスに含める必要がある項目をA対象者の個人情報保護、Bインフォームド・コンセントと倫理審査、C知的財産権の帰属、D寄託する項目/しない項目の判断、Eデータ提供先の制限、の5点に整理した。

一方で、エンドポイントとしての死因情報は非常に価値が高いものであることから、追跡が終了したコホート研究の二次利用を進めるために、死因情報のソースとなる人口動態統計調査の有効活用の方策を模索することが望まれ、その際の論点を2つ提示した。統計法の従来の解釈経緯などといったコホート研究とは異なる議論の方向性が必要と考えられ、医学分野の知見だけではない法学分野からの知見に照らし、今後、より視野を広げた説得力ある議論の展開が不可欠である。なお、がん登録から得たがん罹患情報の研究目的での二次利用についても、今後同様の検討が必要と考えられた。

**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（中央大学理工学部・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ことを目的とした。特に配慮すべき事項の確認のため、先行している国内外の事例収集、実データを用いた一意性の検討、共通化可能項目の検討、アーカイブ化を進めるに際して必要なガイダンス項目の整理を進めた。一方で、統計

法の規定により、人口動態統計情報を利用して把握している死因情報は、データアーカイブに付加して二次利用できないことから、現制度化での運用方法の提案を行うとともに、今後、死因を付加して提供することを可能にするための統計法解釈の論点を提示することを目指した。

B. 研究方法

①米国ならびに国内で先行している社会科学分野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ため、各分野の専門家より情報を得た。

②三府県コホートデータ約10万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。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。

③現制度下で二次的に死因情報を利用するため、研究の都度、コホートデータに死因情報を付与する方法について、JALSデータを用い検討した。また、国内外の医療（臨床）データ連携、疫学共同研究などにおける分散型ネットワークの先行事例を収集した。

④これまでの研究成果や先行する分野のデータアーカイブセンターのガイドライン等を参考に、特に研究開始時期のデータ取り扱いが現在のものと異なり、かつ多数を取り扱う疫学研究データのアーカイブ化の場合にガイダンスに含めておくべき項目を検討した。

⑤統計法の位置づけと人口動態統計資料の利用価値の観点から、情報関係法制に詳しい法律学の専門家である友岡史仁日本大学法学部教授を交え、検討を行った。

⑥がん登録推進法の規定及び同法に関する政省令を読解し、関連する研究者と議論することにより、追跡終了後コホート研究を用いた共通化データベースが構築された場合における、がん登録情報（がん罹患情報と死亡者情報）の利用をめぐる諸問題について考察した。

C. 研究結果

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 社会学分野のデータアーカイブの現状

社会科学系では、特に若手の研究者がデータアーカイブを利用して、オリジナルな枠組みで分析を行い、新たな知見を出していくことがより一般的になってきている。そのためのセンターの1つであるSSJDA（Social Science Japan Data Archive）には現在、約1600件のデータが寄託されており、2013年度は2700件の利用があった。このように活用が進んでいる背景には、データアーカイブセンターが設立されたこと、ならびに二次分析のメリットが広く研究者に認識されたことがある。このようにデータが収集・公開され第三者が分析することは、データの再現性を確認すること

につながる。また、特に公的資金が投入され実施された調査データに関しては、調査者個人のものではないという認識も広まりつつある。

現在、SSJDAの運営費用は文部科学省（2010年度より国立大学法人共同利用・共同研究拠点）、東京大学社会科学研究所から運営費、データアーカイブに関わる科学研究費で賄われている。データアーカイブセンターの活動として行われている業務の主なものは、データ寄託の依頼・受付、データ整理、データ秘匿処理、メタデータの作成、データ利用の受付・提供、リモート集計の提供、二次利用成果の公開、データ寄託者の表彰、二次利用促進と適切な解析のための研究会・セミナーの開催等多岐にわたっている。

②三府県コホートデータによる一意性の検討

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。

さらに三府県コホート対象者約100,000例全てを使用した場合と、無作為抽出により標本数を1万、1,000、100に変化させた場合各々で、分類数$K$とユニークセル数$S_1$、分類数に占めるユニークセル数の割合$S_1 / K$を検討した。全例を使用した場合、分類数$K$が増加するとともに、ユニークセル数$S_1$およびユニークセル数の割合$S_1 / K$は増加するが、ユニークセル数の割合$S_1 / K$については、分類数$K$が約20,000例になるまで急増後80%程度でプラトーに達し、分類数$K$が80,000例あたりからさらに増加する、というパターンを示した。この傾向を全例を使用した場合と標本数を減らした場合で比較すると、分布の形状は似通っていたが、急増する部分の勾配が緩やかになり（100例使用の場合は40例程度まで）、プラトー部分が狭くなる傾向があった。

③ -1　死因情報を外したアーカイブデータを利用する運用案

提案するデータ利用基盤の概略は次のとおりである。①研究コンソーシアムに参加する各研究が、基本データ（生活習慣、検査データなど）と死因を除いた追跡データをアーカイブセンターに提供する。②アーカイブセンターでは、基本データベースと追跡データベースを分けて構築する。その際、基本データベースは原則登録時から修正なしの状態、追跡データベース（その後の死因照合作業で必要となる「死亡地（市町村）」、「死亡日」、「生年月日」、「性別」を含む）は、今後の追跡継続に応じて更新できる構造とする。③このデータベースを利用した研究を行いたい研究者は、死因情報を得るために厚生労働省に対し人口動態調査二次利用申請を行う。④承認後に提供を受けた死因情報をアーカイブセンター内で、「死亡地（市町村）」、「死亡日」、「生年月日」、「性別」をキー変数として、保有する追跡情報と照合する。⑤死因を付与した一時的な解析データセットを作成し、研究計画に基づいた解析に使用する。⑥研究終了後は死因情報を削除(抹消)し、厚生労働省に利用後報告を行う。

この方法の妥当性を確認するため、JALS対象地域の市町村で1999年1月1日から2012年12月31日までに発生した死亡の調査票情報を厚生労働省に申請、入手した。JALSの対象者で、当該期間中に死亡が特定できていた7,137件（職域コホートと死亡調査データが確定していないコホートを除く）のうち、99.5%が、性別、生年月日、死亡年月日、死亡時の居住市町村名をキー変数として人口動態調査データと一致した。なお、不一致のうち14件は以前にJALSが行った死因照合作業において既に未照合が判明しており、各コホートに対して死亡時情報を確認したがいずれも情報に誤りがなく、人口動態統計作成の過程で入力間違い

等が発生した事例と判断した。このため、今回照合出来なかった例は、実質として19件（0.27%）であった。

③ -2　分散型ネットワーク国内外の医療（臨床）データ連携、疫学共同研究利用などの先行事例収集と、疫学追跡終了後コホートデータの共通利用についての検討

近年整備されつつある分散型ネットワークについて、医療（臨床）データ連携、国際疫学共同研究での5つの先行事例を収集し検討を行ったところ、医療（臨床）と疫学研究の事例では、データ連携(交換)の目的(必要とする背景)、運用方法などが異なっていることが明らかとなった。

④アーカイブデータ提供のためのガイダンス項目

含める必要があると考えられたのは、以下の項目であった。

A　対象者の個人情報保護：実データを用いた我々の過去2年間の検討では調査項目が多岐に渡り、かつ対象者数が多いコホート研究では、データの一意性はあるものと考えた方がよいと結論した。回答者のプライバシー保護の観点から、調査地域を粗く束ねるなどの手段を用いて、個々の回答者の識別を不可能にすることが望まれる。

B　インフォームド・コンセントと倫理審査：新たに開始される研究では、対象者からデータ寄託に関する同意が得られていることを原則とした上で、寄託者の所属機関で倫理審査を受けることが可能である。しかし、既に追跡が終わったコホート研究の場合、その開始は今から20年以上前であると考えられ、研究の説明同意プロセスは現在と異なる。そのため、対象者の同意あるいは再同意の取得を前提とすることは困難であり、所属機関での倫理審査に加え何らかの形で二次利用に関する広報・情報公開を行うことが適切と考えられた。

C　知的財産権の帰属：生体試料を伴わない情報のみのアーカイブでは、知的財産権の発生は稀と思われるが、生体試料を収集・保管するコホート研究では、研究終了時にも試料が残ることが想定される。二次利用者とデータ寄託者の間の分配方法については、一定のルールが定まっていないことから、利用開始時に取り決めを交わしておくことが望ましい。

D　寄託する項目/しない項目の判断：個人同定の可能性が特に高い項目は、丸め処理をするか、提供しないことが望まれる。また、調査方法や設定が特殊である等何らかの理由により、研究内容を十分に理解しない二次利用者の解析・解釈によって誤った結論を引き出すおそれがある項目についても、提供しないことが推奨される。

Eデータ提供先の制限：二次利用者の資格、利用上の制約等については、各研究関係者の意向を反映するような区分けの周知とそれに沿った対応が求められると考えられた。

⑤統計法解釈に関する検討

コホート研究に不可欠とされる死因情報を人口動態統計資料から得ることは、統計法上の目的外利用を意味するため、その活用に当たり、統計法上設けられた諸手続といった一定の制約が生じている。しかし、データの二次利用により、公費を投入し多くの人々の協力を得て収集された試料・情報の有効利用、若手研究者の育成、データの検証等幅広い研究利用とその成果に直結し、ひいてはより正確な健康関連要因の解明に具体的に貢献しうることを考え合わせれば、統計法の規定による制約を受けない形で、死因も含めたより広い形による該当データの利用可能性が切望される。このことから、当該情報を長期的かつ利用者の範囲を広げ、コホート研究に有効活用する道を法制度について論ずる意義は、十分にあると思われる。

この場合の論点として、A統計法の解釈運用を変更することで足りるのか、B統計法とは異なる別立法を行う必要性があるのか、という二点が考えられる。Aの場合は、上記の制度趣旨と過去の運用の経緯に照らして乗り越えるべき解釈論を構築する必要があると考えられ、その意味で、死因情報以外の件についても考慮に入れた制度全体の観点から議論を要するものと思われる。対してB

の場合は、Aとは異なり立法的措置により解決するため、特別法として位置付けられ、法解釈上の齟齬はきたさないと思われる一方、既存の仕組みとは異なる新たな制度設計を求められるぶん、個人情報保護に照らした保護に係る必要なスタンダードを充足する詳細な議論を要する。AまたはBのいずれを選択するにせよ、新たな立法の煩雑さや既存の仕組みの有効活用という効率性の視点からすれば、現時点ではAの選択肢が方向性として現実的と考えられた。

⑥がん登録推進法における死亡者情報票の活用

病院・診療所と市町村から届けられた情報をもとに全国がん登録データベースが構築される。同データベースには、がん罹患情報と罹患者における生存死亡に関する情報の２つが含まれる。そして、生存死亡の確認の期間は、政令により百年間と定められた。

がん登録情報をコホート研究に活用する際は、匿名化されない情報の提供を受ける必要がある。そのための要件の１つが調査対象者の同意であり、「提供の求めを受けた情報に係るがん罹患者が生存している場合、その調査研究を行う者が、当該がん罹患者から調査研究目的で当該がん登録情報が提供されることについて同意を得ていること」とされている。

ただし、がん登録推進法施行日（平成28年１月１日）以前に開始されたがんに係る調査研究ついては、同法附則２条（経過措置）により「同意を得ることが当該がんに係る調査研究の円滑な遂行に支障を及ぼすものと認められる場合として政令で定める場合に該当するものである場合において、（略）これらの同意に代わる措置として厚生労働大臣が定める指針に従った措置が講じられているときは、（略）適用しない」とされている。

「同意を得ることが当該がんに係る調査研究の円滑な遂行に支障を及ぼすものと認められる場合」については、がん登録等の推進に関する法律施行令（平成27年９月９日政令323号）により以下の各号のいずれかに該当する場合と定められた。

一　調査研究の対象者が五千人以上の場合

二　調査対象者と連絡を取ることが困難であること、または対象者の同意を得ることが調査研究の結果に影響を与えること。

そして「同意に代わる措置として厚生労働大臣が定める指針」が、厚生労働省告示第471号（平成27年12月15日）によって定められた。それによると、(1)「人を対象とする医学系研究に関する倫理指針」第５章に即して研究対象者等からインフォームド・コンセントの取得等を実施していること、(2) 全国がん登録情報等の提供を受けることについて情報公開等の措置を講ずること、(3) 研究対象者が提供について拒否できる機会を保障すること、という３点が同意代替措置とされた。

D. 考察

国民の税金を投入し、多くの人手と長期の追跡を経て構築されたコホート研究データをアーカイブ化し、広く二次利用を可能とすることは、研究の透明性確保、第三者による研究結果の検証、若手の育成に寄与するのみならず、研究の無駄・重複を減らし、必要な公費・労力を新しい有意義な研究に向けるという意義もある。

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。また、国内でも社会科学系分野は先行しており、そのためのセンターが設立され、現在では多くのデータが二次利用されている。しかし、今の形になるまでに、10～20年の年月を要しており、データ寄託がある一定数に達するまで、利用のメリットが十分に浸透するよう働きかけるとともに、利用のための環境整備も必要と考えられた。

前述したようにアーカイブ化により二次利用環境を整える意義は大きいが、疫学研究では生活習慣や生活環境のみならず、個人の心身面の健康状態情報を収集することから、その取り扱いには慎重であることが望まれる。特に三府県コホートを用いた一意性の検討からも明らかなように、多数を対象に多項目のデータを長期にわたって収集するコホート研究では、一意性があるものとの前提で対応を考える必要がある。今回、特に追跡の終わったコホート研究の利活用を進めることを念頭に、データ提供の際に従うべきガイダンスに含める必要がある項目を5点に整理した。今後、追跡が終了したコホート研究データの寄託を進める際の指標になるものと考えられる。

多くのコホート研究では、死因情報を人口動態統計資料から得ており、統計法の規定上、現状では非常に重要な情報であるにもかかわらずアーカイブデータに死因情報を付加して二次利用に供することはできない。そこで、現制度下での運用方法につき、2つの提案を行った。提案1の運用案を複数のコホート研究から構成される国内循環器疫学研究（JALS）に適用したところ、死因照合率は99.7%であり、照合作業の技術的側面、作業手順化の面で問題はなく、運用案の一つになりうると考えられた。提案2については、医療（臨床）、疫学研究で先行している分散型ネットワークの事例を収集し、本検討事例である疫学追跡終了後コホートデータの共通利用の場面に応用可能な事例があるか検討を行った。疫学共同研究における分散型ネットワークを用いたデータ連携の例などから、リモート環境で様々な統計解析を柔軟に行えること、ローカルサイト（大学、研究所、医療機関を想定）へアクセスを行う上でのセキュリティ、さらには、人口動態統計の二次利用をローカルサイト単位で各々申請し、死因を付与したデータを統合して用いることや、施設外部のPC上のRAMへの一時的な書き出し（保存はされない、電子データの「一時的蓄積」）の可否など、法令解釈の問題にも留意する必要があると考えられた。疫学研究追跡終了後のコホートデータの共通利用に関して、提案した二つの方法で運用していくことは可能と考えられるが、実際に運用する場合には、前者においてはデータ利用の規約、データ利用の物理的環境、サポート体制、後者は、人口動態二次利用申請による死因データ利用の法令解釈などの課題があると考えられた。

一方でコホート研究にとって死因情報は非常に重要なものであり、利用できないことによりアーカイブデータの二次利用の価値は低下すると考えられる。また、米国ではNational Death Index（NDI）という、厚生省（U.S. Department of Health and Human Services）の下部機関が、研究目的での生存・死亡確認情報（死亡時には、死亡年月日や死因などを含む）の提供を行っている。研究者は、調査対象者リスト（氏名、性、生年月日、住所、社会保障番号など）を提出し、審査にパスすると、有料（基本料350ドル＋対象者1人1年あたり15セント）で、上記情報が提供される。これにより、米国の疫学研究・臨床研究のレベルと即時性は飛躍的に向上し、医学研究や医薬品開発において国際的に有利な地位を確保することができたといえ、今のままでは日本の疫学研究は後塵を拝する。そこで、法の解釈、または立法により、コホート研究において重要な追跡情報となる死因を公的調査情報（人口動態統計資料）から得た上で二次的利用を可能にできないかを検討した。しかし、これには統計法の従来の解釈経緯などといった従前のコホート研究とは異なる方向性が必要とされており、今後も継続して、法学分野からの知見に照らし、より視野を広げた説得力ある議論を展開する必要があると考えられた。また、がん登録推進法の施行により、追跡情報とし

てがん罹患情報の利用も広がるものと考えられる。しかし、がん登録推進法と同法に関する政省令の規定において定められた要件・方法などに従って、がん登録情報の研究利用の承認を受けた者が、データアーカイブにより情報を第三者に研究目的で提供することの可否については、それらの規定の中では一切言及されていない。この問題については、コホート研究で人口動態統計資料から得た死因情報をデータアーカイブに付加して提供するための法解釈と同様の検討が必要になると思われる。

## E. 結論

疫学研究により得られたデータを広く共有化するためのシステムであるデータアーカイブ化に向けた検討を行った。統計法の規定上、死因情報を付加してのデータ公開・二次利用は認められていないことなどを踏まえて、現制度下での運用方法を2つ提案した。一つ目はアーカイブセンターなどに死因情報以外のデータを集約管理し、必要に応じて従来通りの死因照合作業を実施し、死因を付加したデータセットを作成し解析を行う方法である。二つ目は、データを保持者のもとに置いたまま、必要と判断した情報だけを選択的に共有させる分散型ネットワークによる方法である。また、実データを用いて、一意性に関する検討を行い、10万人規模のコホート集団はもちろん、100例規模のコホート集団でも分類数が全対象者数の40%程度で、ユニークセルの割合が80%に達していたことから、コホートの規模にかかわらず、80%程度のレコードは一意性があるものとして対応する必要があることを示した。これら研究成果や他国・他領域での実例を参考に、データアーカイブの利活用を進めるため、データ提供の際に従うべきガイダンスに含める必要がある項目をA対象者の個人情報保護、Bインフォームド・コンセントと倫理審査、C知的財産権の帰属、D寄託する項目/しない項目の判断、Eデータ提供先の制限、の5点に整理した。

一方で、エンドポイントとしての死因情報は非常に価値が高いものであることから、追跡が終了したコホート研究の二次利用を進めるために、死因情報のソースとなる人口動態統計調査の有効活用の方策を模索することが望まれ、その際の論点を2つ提示した。統計法の従来の解釈経緯などといったコホート研究とは異なる議論の方向性が必要と考えられ、医学分野の知見だけではない法学分野からの知見に照らし、今後、より視野を広げた説得力ある議論の展開が不可欠である。なお、がん登録から得たがん罹患情報の研究目的での二次利用についても、今後同様の検討が必要と考えられた。

## 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

# 分担報告書

厚生労働省科学研究費補助金(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」
分担研究報告書

# 他国・他領域におけるデータアーカイブ利活用環境
## —米国におけるデータアーカイブの研究利用と国内社会科学分野のデータアーカイブの現状—

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**

疫学研究により得られたデータを広く共有化するためのシステムであるデータアーカイブ化に向けた課題を整理するために、先行している米国、国内社会科学分野の現状を把握し、疫学研究、とくに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の知見を得た。

**A. 研究目的**

データアーカイブを二次利用するメリットは、

1）既に行われている調査を繰り返さずに済み労力、資金とも無駄な投入を避けることができること、特に多くの変数を得るような調査では得られたすべての情報を調査者が解析することはないため、利用されていない変数について独自のアイディアで解析することで、新たな知見を得ることができること、
2）若手研究者にとっては、自身で小規模な回収率の高くない調査を行うことに比べ、質のよい調査データにアクセスできること、
3）学生教育の際にも、実データを用いた教育を行うことができること、　である。

しかし、疫学研究のアーカイブデータの提供・利用環境はまだ整っていない。そこで、米国ならびに国内で先行している社会科学分野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. 研究方法**

①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。
②社会科学系分野のデータアーカイブセンターの一つ、東京大学社会科学研究所の附属社会調査・データアーカイブ研究センターの藤原翔氏より、情報提供をいただいた。

**C. 研究結果**

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、コホート研究を一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、倫理審査を事前に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。

②社会学分野のデータアーカイブの現状

社会科学系では、特に若手の研究者がデータアーカイブを利用して、オリジナルな枠組みで分析を行い、新たな知見を出していくことがより一般的になってきている。そのためのセンターのひとつであるSSJDA（Social Science Japan Data Archive）には現在、約1600件のデータが寄託されており、2013年度は2700件の利用があった。このように活用が進んでいる背景には、データアーカイブセンターが設立されたこと、ならびに二次分析のメリットが広く研究者に認識されたことがある。このようにデータが収集・公開され第三者が分析することは、データ解析の再現性を確認することにつながる。また、特に公的資金が投入され実施された調査データに関しては、調査者個人のものではないという認識も広まりつつある。

現在、SSJDAの運営費用は文部科学省（2010年度より国立大学法人共同利用・共同研究拠点）、東京大学社会科学研究所から運営費、データアーカイブに関わる科学研究費で賄われている。データアーカイブセンターの活動として行われている業務の主なものは、データ寄託の依頼・受付、データ整理、データ秘匿処理、メタデータの作成、データ利用の受付・提供、リモート集計の提供、二次利用成果の公開、データ寄託者の表彰、二次利用促進と適切な解析のための研究会・セミナーの開催等、多岐にわたっている。

## D. 考察

国民の税金を投入し、多くの人手と長期の追跡を経て構築されたコホート研究データをアーカイブ化、広く利用可能にすることは、研究の透明性確保、第三者による研究結果の検証、若手の育成に寄与するのみならず、研究の無駄・重複を減らし、必要な公費・労力を新しい有意義な研究に向けるという意義もある。そこで、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・単純化と必要経費の設定、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。

また、国内でも社会科学系分野のデータアーカイブは先行しており、そのためのセンターが設立され、現在では多くのデータが二次利用されている。しかし、今の形になるまでに、10～20年の年月を要しており、データ寄託がある一定数に達するまで、利用のメリットが十分に浸透するよう働きかけるとともに、利用のための環境整備も必要と考えられた。

## E. 結論

疫学研究により得られたデータを広く共有化す

るためのシステムであるデータアーカイブ化に向けた課題を整理するために、先行している米国、国内社会科学分野の現状を把握し、疫学研究、とくに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の知見を得た。

**F. 健康機器情報**

特になし

**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

厚生労働省科学研究費補助金(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」
分担研究報告書

# 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。また、標本数を変化させた場合にそれぞれどのような頻度で一意性が見られるかを比較した。100,629例全てを使用した場合と標本数を減らした場合、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が増加するという分布の形状は似通っていたが、標本数が大きい場合ほど少ない分類数でユニークセルの割合が高率に達していた。コホートの規模にかかわらず、80%程度のレコードは一意性があった。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. 研究方法

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,…,N)とする。つまり、$\Sigma S_i=K$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。

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①アンケートカテゴリのみ、②【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、③【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、④【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。

さらに、三府県コホートデータを使用し 100,629 例全てを使用した場合と、無作為抽出により標本数を 1 万、1,000、100 に変化させた場合について検討を行う。分析対象となる変数は、前述の解析と同様、226 変数からなる個人レコードのうち、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 変数とした。
[検討 5]
100,629 例全てを使用した場合と、無作為抽出により標本数を 1 万、1,000、100 に変化させた場合について、21 のアンケートカテゴリに対し①アンケートカテゴリのみ、②【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、③【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、④【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］①~④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。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。

［検討5］100,629例全てを使用した場合と、無作為抽出により標本数を1万、1,000、100に変化させた場合について、①~④の組み合わせから得られた461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2）。

また100,629例全てを使用した場合と、無作為抽出により標本数を1万、1,000、100に変化させた場合について、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3）。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4）。100,629例全てを使用した場合、分類数Kが増加するとともに、ユニークセル数S_1およびユニークセル数の割合S_1/Kは増加するが、ユニークセル数の割合S_1/Kについては、分類数Kが約20,000例になるまで急増し、次に80%程度でプラトーに達し、分類数Kが80,000例あたりからさらに増加する、というパターンを示した。100,629例全てを使用した場合と標本数を減らした場合を比較すると、分布の形状は似通っていたが、急増する部分の勾配が緩やかになり（100例使用の場合は40例程度まで）、プラトーに達する部分が狭くなる傾向があった。

## 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よう、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。

10万人規模のコホート集団の場合、分類数が全対象者数の概ね20,000程度で、ユニークセルの割合が80%に達していた。対象者数を少なくするにつれて、立ち上がりが緩やかになり、100例規模のコホート集団では、分類数が40程度で、ユニークセルの割合が80%に達していた。コホートの規

模にかかわらず、80%程度のレコードは一意性があるものとして対応する必要がある。

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。

10万人規模のコホート集団の場合、分類数が全対象者数の20%程度で、ユニークセルの割合が80%に達していた。100例規模のコホート集団では、分類数が全対象者数の40%程度で、ユニークセルの割合が80%に達していた。コホートの規模にかかわらず、80%程度のレコードは一意性があるものとして対応する必要がある。

E. 健康機器情報

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1　対象となる変数について（グレーの変数は解析対象外）

| 変数名 | 属性 | 内容 | 定義の相違 | 凡例など | 備考 |
|---|---|---|---|---|---|
| v0000 | 数値 | 通しID | 共通 | 1-100629 | 府県(v0400)、自治体(v0403)、地域固有ID(v0001)のソート順 |
| v0001 | 数値 | 地域固有ID | 府県間で異なる | (欠損なし) | |
| v0100 | 数値 | 観察期間=負　フラグ | 共通 | 1: Yes(欠損=.) | 解析時除外。<br>追跡開始日(v0500)>観察終了日(v0521)。<br>愛知都市地区のみ14例。 |
| v0200 | 数値 | 性別 | 共通 | 0:男, 1:女(欠損なし) | |
| v0300 | 数値 | 追跡開始年齢 | 共通 | 40-99(欠損なし) | |
| v0400 | 数値 | 府県番号 | 共通 | 1:宮城, 2:愛知, 3:大阪 | |
| v0401 | 数値 | 地区番号 | 共通 | 1: 都市地区, 0: 対照地区 | |
| v0402 | 数値 | 地区通し番号 | 共通 | 1: 宮城対照, 2: 宮城都市, 3: 愛知対照, 4: 愛知都市, 5: 大阪対照, 6: 大阪都市 | |
| v0403 | 数値 | 自治体番号 | 府県間で異なる | 宮城:　1:仙台,2:涌谷,3:田尻;　愛知:　1:名古屋,2:犬山;　大阪:　1:東成,2:能勢,3:河南,4:熊取 | |
| v0500 | SAS日付 | 追跡開始日 | 共通 | (欠損なし) | |
| v0501 | SAS日付 | 10年コホート追跡終了日(地域ごとの定義) | 共通(ただし日付は地区間で異なる) | 宮城:　1993/12/31, 愛知都市地区:　1995/10/31, 愛知対照地区:　1995/6/30, 大阪東成:　1994/10/31, 大阪能勢:　1993/1/31, 大阪河南:　1994/11/30, 大阪熊取:　1995/1/31 | |
| v0502 | SAS日付 | 10年観察終了日(個人ごとの定義) | 共通 | コホート10年追跡終了日, 転出日, 死亡日のうち一番早い日(欠損なし) | |
| v0503 | 数値 | 10年転帰 | 共通 | 3: コホート追跡終了日において生存, 6: コホート追跡終了日までに死亡, 9: コホート追跡終了日までに転出(欠損なし) | |
| v0511 | SAS日付 | 15年コホート追跡終了日(地域ごとの定義) | 共通(ただし日付は地区間で異なる) | 宮城:　1993/12/31, 愛知都市地区:　2000/10/31, 愛知対照地区:　2000/6/30, 大阪東成:　1999/10/31, 大阪能勢:　1998/1/31, 大阪河南:　1999/11/30, 大阪熊取:　2000/1/31(欠損なし) | |
| v0512 | SAS日付 | 15年観察終了日(個人ごとの定義) | 共通 | コホート15年追跡終了日, 転出日, 死亡日のうち一番早い日(欠損なし) | |
| v0513 | 数値 | 15年転帰 | 共通 | 3: コホート追跡終了日において生存, 6: コホート追跡終了日までに死亡, 9: コホート追跡終了日までに転出(欠損なし) | |
| v0520 | 数値 | 転帰(受領データママ) | 共通 | 3: 生存, 6: 死亡, 9: 転出(欠損なし) | 各地域の最終確認時の転帰。解析時は追跡終了日定義の上調整が必要。<br>愛知は転出(v0520=9)で死因ありのデータが都市地区277例、対照地区73例あり→転出として扱い、死因は無視(若井先生確認済)。 |
| v0521 | SAS日付 | 観察終了日(個人ごとの定義:　受領データママ) | 共通 | コホート追跡終了日による定義をする前の個人ごとの観察終了日。 | 愛知のみコホート追跡終了日より後の日付あり(2000/12/31まで)。 |
| v0600 | 数値 | 死因ICD-9コード4桁 | 共通 | ICD-9(例:　38.9, 150.1) | 基本となる死因変数(ただし、宮城の15年死亡はv0601,v0602とOR結合が必要)。<br>10年死亡の死因別死亡例はこの変数のみで同定可。 |
| v0601 | 数値 | 死因ICD-9コード4桁(宮城15年死亡) | 宮城のみ | ICD-9(例:　38.9, 150.1) | 宮城15年死亡の死因はv0600(ICD-9), v0601(ICD-9), v0602(ICD-10)の3変数に入っている(これら3変数間に重複なし)。<br>15年死亡の死因別死亡同定にはこれら3変数のOR結合が必要。 |
| v0602 | 文字 | 死因ICD-10コード4桁(宮城15年死亡) | 宮城のみ | ICD-10(例:　C030, I219) | |
| v0610 | 文字 | 死因ICD-9コード | 府県間で異なる | ICD-9(元データのママのため、形式が異なる)。 | v0600の数値化前データ |
| v0611 | 文字 | 死因ICD-9またはICD-10コード(宮城15年死亡) | 宮城のみ | ICD-9とICD-10が混在(元データのママ) | v0601, v0602のICD-9とICD-10に分ける前のデータ |
| v0620 | 数値 | 10年死因不明フラグ | 共通 | 1: Yes(欠損=.) | v0600=.　かつ　v0611=""　かつ 10年死亡(v0503=6)。<br>10年死亡例中19例あり(すべて宮城)。 |
| v0621 | 数値 | 15年死因不明フラグ | 共通 | 1: Yes(欠損=.) | v0600=. かつ v0611="" かつ15年死亡(v0513=6)。<br>15年死亡例中29例あり(宮城28例、大阪東成1例)。 |
| v1000 | 数値 | 自覚症状あるか　咳 | 共通 | 1: Yes(欠損=.) | 明示回答ありなら1、それ以外は欠損。<br>宮城、大阪熊取はv1011と矛盾ありママ。 |
| v1001 | 数値 | 自覚症状あるか　痰 | 共通 | 1: Yes(欠損=.) | 明示回答ありなら1、それ以外は欠損。<br>宮城、大阪熊取はv1011と矛盾ありママ。 |
| v1002 | 数値 | 自覚症状あるか　血痰 | 共通 | 1: Yes(欠損=.) | 明示回答ありなら1、それ以外は欠損。<br>宮城、大阪熊取はv1011と矛盾ありママ。 |
| v1003 | 数値 | 自覚症状あるか　動悸 | 共通 | 1: Yes(欠損=.) | 明示回答ありなら1、それ以外は欠損。<br>宮城、大阪熊取はv1011と矛盾ありママ。 |
| v1004 | 数値 | 自覚症状あるか　息切れ | 共通 | 1: Yes(欠損=.) | 大阪東成は欠損。<br>明示回答ありなら1、それ以外は欠損。<br>宮城、大阪熊取はv1011と矛盾ありママ。 |
| v1005 | 数値 | 自覚症状あるか　食欲不振 | 共通 | 1: Yes(欠損=.) | 大阪東成は欠損。<br>明示回答ありなら1、それ以外は欠損。<br>宮城、大阪熊取はv1011と矛盾ありママ。 |
| v1006 | 数値 | 自覚症状あるか　便秘 | 共通 | 1: Yes(欠損=.) | 大阪東成は欠損。<br>明示回答ありなら1、それ以外は欠損。<br>宮城、大阪熊取はv1011と矛盾ありママ。 |
| v1007 | 数値 | 自覚症状あるか　不眠 | 共通 | 1: Yes(欠損=.) | 愛知、大阪熊取、大阪河南は欠損。<br>明示回答ありなら1、それ以外は欠損。 |
| v1008 | 数値 | 自覚症状あるか　やせてきた | 共通 | 1: Yes(欠損=.) | 明示回答ありなら1、それ以外は欠損。<br>宮城、大阪熊取はv1011と矛盾ありママ。 |
| v1009 | 数値 | 自覚症状あるか　疲れやすい | 共通 | 1: Yes(欠損=.) | 明示回答ありなら1、それ以外は欠損。<br>宮城、大阪熊取はv1011と矛盾ありママ。 |
| v1010 | 数値 | 自覚症状あるか　喘息の発作 | 共通 | 1: Yes(欠損=.) | 宮城、大阪東成は欠損。<br>明示回答ありなら1、それ以外は欠損。 |
| v1011 | 数値 | 自覚症状あるか　なし | 共通 | 1: Yes (欠損=.) | 「なし」回答ありの場合1、それ以外は欠損。<br>宮城、大阪熊取はv1000-v1009と矛盾ありママ。 |
| v1100 | 数値 | 既往　高血圧 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし(欠損=.) | 宮城、大阪東成、大阪河南は疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。<br>宮城、大阪熊取はv1120との矛盾ママ。 |
| v1101 | 数値 | 既往　心臓病 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし(欠損=.) | 宮城、大阪東成、大阪河南は疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。<br>宮城、大阪熊取はv1120との矛盾ママ。 |
| v1102 | 数値 | 既往　脳卒中 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし(欠損=.) | 宮城、大阪東成、大阪河南は疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。<br>宮城、大阪熊取はv1120との矛盾ママ。 |
| v1103 | 数値 | 既往　糖尿病 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし(欠損=.) | 宮城、大阪東成、大阪河南は疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。<br>宮城、大阪熊取はv1120との矛盾ママ。 |
| v1104 | 数値 | 既往　肝臓病 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし(欠損=.) | 宮城、大阪東成、大阪河南は疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。<br>宮城、大阪熊取はv1120との矛盾ママ。 |
| v1105 | 数値 | 既往　腎臓病 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし(欠損=.) | 宮城、大阪東成、大阪河南は欠損。 |
| v1106 | 数値 | 既往　胃十二指腸潰瘍 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし(欠損=.) | 大阪能勢は欠損。<br>宮城、大阪東成、大阪河南は疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。<br>宮城、大阪熊取はv1120との矛盾ママ。 |
| v1107 | 数値 | 既往　胆石 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし(欠損=.) | 大阪東成、大阪能勢、大阪河南は欠損。<br>宮城、大阪河南は疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。<br>宮城、大阪熊取はv1120との矛盾ママ。 |

| 変数名 | 属性 | 内容 | 定義の相違 | 凡例など | 備考 |
|---|---|---|---|---|---|
| v1108 | 数値 | 既往 虫垂炎 | 一部相違 | 1: Yes (欠損=.) | 宮城のみ。<br>疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。 |
| v1109 | 数値 | 既往 結核ろく膜炎 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし (欠損=.) | 宮城、大阪東成、大阪河南は疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。<br>大阪能勢は結核と肋膜炎の結合（選択肢1,2は1優先で結核と肋膜炎のOR結合、,選択肢3は1,2でない場合に結核と肋膜炎のAND結合）。<br>宮城、大阪熊取はv1120との矛盾ママ。 |
| v1110 | 数値 | 既往 肺炎 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし (欠損=.) | 宮城、大阪東成、大阪河南は疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。<br>宮城、大阪熊取はv1120との矛盾ママ。 |
| v1111 | 数値 | 既往 喘息 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし (欠損=.) | 宮城、大阪東成、大阪河南は疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。<br>宮城、大阪熊取はv1120との矛盾ママ。 |
| v1112 | 数値 | 既往 慢性気管支炎 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし (欠損=.) | 宮城、大阪東成、大阪河南は疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。<br>宮城、大阪熊取はv1120との矛盾ママ。 |
| v1113 | 数値 | 既往 肺気腫 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし (欠損=.) | 宮城、大阪東成、大阪河南は疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。<br>宮城、大阪熊取はv1120との矛盾ママ。 |
| v1114 | 数値 | 既往 じん肺 | 一部相違 | 1: 現在あり, 2: 以前あった, 3: 今までなし (欠損=.) | 宮城、大阪東成、大阪河南は疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。<br>宮城、大阪熊取はv1120との矛盾ママ。 |
| v1115 | 数値 | 既往 貧血 | 共通 | 1: Yes (欠損=.) | 大阪東成、大阪河南のみ。<br>疾病番号選択のみのため、回答ありを便宜上1に割り振り回答なしは欠損。 |
| v1116 | 数値 | 既往 アレルギー性鼻炎 | 共通 | 1: 現在あり, 2: 以前あった, 3: 今までなし (欠損=.) | 大阪能勢のみ。 |
| v1117 | 数値 | 既往 蕁麻疹 | 共通 | 1: 現在あり, 2: 以前あった, 3: 今までなし (欠損=.) | 大阪能勢のみ。 |
| v1118 | 数値 | 既往 湿疹・皮膚炎 | 共通 | 1: 現在あり, 2: 以前あった, 3: 今までなし (欠損=.) | 大阪能勢のみ。 |
| v1119 | 数値 | 既往 痛風 | 共通 | 1: Yes (欠損=.) | 宮城のみ |
| v1120 | 数値 | 既往 なし | 共通 | 1: Yes (欠損=.) | 「なし」回答ありの場合1、それ以外は欠損。<br>大阪能勢欠損。<br>宮城、大阪熊取はv1100-v1119との矛盾ママ。 |
| v1200 | 数値 | 身長(cm) | 共通 | (欠損=.) | 外れ値ママ |
| v1201 | 数値 | 体重(kg) | 共通 | (欠損=.) | 外れ値ママ |
| v1202 | 数値 | BMI連続変数 | 共通 | (欠損=.) | v1201/v1200/v1200*10000で計算。v1200またはv1201が0, 999.9, 欠損のいずれかの場合は欠損。 |
| v1300 | 数値 | 胸部レントゲン検査受診有無 | 一部相違 | 1: Yes, 0: No (欠損=.) | 宮城、愛知、大阪熊取は「胸部のレントゲン検査(結核検診)」。<br>宮城は「昨年(昭和58年)1年間に」、大阪東成、大阪河南は「この1年のあいだに」、それ以外は単に「受けたことがありますか」と聞いて下位で時期質問。 |
| v1301 | 数値 | 胸部レントゲン検査受診有無(1年以内) | 一部相違 | 1: Yes (欠損=.) | 宮城、大阪東成、大阪河南はv1300=1、愛知はv1300=1かつ昭和60年受診、大阪能勢はv1300=1かつ昭和57年度受診、大阪熊取はv1300=1かつ昭和59年受診。<br>注： ベースライン調査は、宮城は昭和59年1～2月、愛知名古屋は昭和60年10～11月、愛知犬山は昭和60年7～8月、大阪能勢は昭和58年2月、大阪熊取は昭和60年2～3月、大阪河南は昭和59年11～12月、大阪東成は昭和59年10～11月に実施。 |
| v1302 | 数値 | 胸部レントゲン検査受診有無(1年より前～2年以内) | 一部相違 | 1: Yes (欠損=.) | 宮城、大阪東成、大阪河南は欠損。愛知はv1300=1かつ昭和59年受診、大阪能勢はv1300=1かつ昭和56年度受診、大阪熊取はv1300=1かつ昭和58年受診。<br>注： ベースライン調査は、宮城は昭和59年1～2月、愛知名古屋は昭和60年10～11月、愛知犬山は昭和60年7～8月、大阪能勢は昭和58年2月、大阪熊取は昭和60年2～3月、大阪河南は昭和59年11～12月、大阪東成は昭和59年10～11月に実施。 |
| v1303 | 数値 | 胸部レントゲン検査受診有無(2年より前) | 一部相違 | 1: Yes (欠損=.) | 宮城、大阪東成、大阪河南は欠損。愛知はv1300=1かつ昭和59年より前受診、大阪能勢はv1300=1かつ昭和56年度より前受診、大阪熊取はv1300=1かつ昭和58年より前受診。<br>注： ベースライン調査は、宮城は昭和59年1～2月、愛知名古屋は昭和60年10～11月、愛知犬山は昭和60年7～8月、大阪能勢は昭和58年2月、大阪熊取は昭和60年2～3月、大阪河南は昭和59年11～12月、大阪東成は昭和59年10～11月に実施。 |
| v1304 | 数値 | 胸部レントゲン検査受診場所(住民健診) | 一部相違 | 1: Yes (欠損=.) | 大阪熊取は「集団検診で(例えば成人病予防協会検診車)」、大阪東成は欠損。<br>v1300=1かつ当該場所の受診有無・受診年いずれか回答ありなら1。 |
| v1305 | 数値 | 胸部レントゲン検査受診場所(保健所・保健センター) | 一部相違 | 1: Yes (欠損=.) | 大阪東成は「保健所や小学校の市民検診」。<br>v1300=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1306 | 数値 | 胸部レントゲン検査受診場所(職場) | 一部相違 | 1: Yes (欠損=.) | v1300=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1307 | 数値 | 胸部レントゲン検査受診場所(病院など) | 一部相違 | 1: Yes (欠損=.) | v1300=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1308 | 数値 | 胸部レントゲン検査受診場所(農協) | 一部相違 | 1: Yes (欠損=.) | 大阪熊取と大阪河南のみ。<br>v1300=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1309 | 数値 | 胸部レントゲン検査受診場所(その他) | 一部相違 | 1: Yes (欠損=.) | v1300=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1400 | 数値 | 胃レントゲン検査受診有無 | 一部相違 | 1: Yes, 0: No (欠損=.) | 宮城、愛知対照、大阪熊取は「胃のレントゲン検査(胃腸病検診)」; 大阪東成、大阪河南は「胃(ガン)検診」。<br>宮城は「昨年(昭和58年)1年間に」、大阪東成、大阪河南は「この1年のあいだに」、それ以外は単に「受けたことがありますか」と聞いて下位で時期質問。 |
| v1401 | 数値 | 胃レントゲン検査受診有無(1年以内) | 一部相違 | 1: Yes (欠損=.) | 宮城、大阪東成、大阪河南はv1400=1、愛知はv1400=1かつ昭和60年受診、大阪能勢はv1400=1かつ昭和57年度受診、大阪熊取はv1400=1かつ昭和59年受診。<br>注： ベースライン調査は、宮城は昭和59年1～2月、愛知名古屋は昭和60年10～11月、愛知犬山は昭和60年7～8月、大阪能勢は昭和58年2月、大阪熊取は昭和60年2～3月、大阪河南は昭和59年11～12月、大阪東成は昭和59年10～11月に実施。 |

| 変数名 | 属性 | 内容 | 定義の相違 | 凡例など | 備考 |
|---|---|---|---|---|---|
| v1402 | 数値 | 胃レントゲン検査受診有無(1年より前～2年以内) | 一部相違 | 1: Yes (欠損=.) | 宮城、大阪東成、大阪河南は欠損。愛知はv1400=1かつ昭和59年受診、大阪能勢はv1400=1かつ昭和56年度受診、大阪熊取はv1400=1かつ昭和58年受診。<br>注: ベースライン調査は、宮城は昭和59年1～2月、愛知名古屋は昭和60年10～11月、愛知犬山は昭和60年7～8月、大阪能勢は昭和58年2月、大阪熊取は昭和60年2～3月、大阪河南は昭和59年11～12月、大阪東成は昭和59年10～11月に実施。 |
| v1403 | 数値 | 胃レントゲン検査受診有無(2年より前) | 一部相違 | 1: Yes (欠損=.) | 宮城、大阪東成、大阪河南は欠損。愛知はv1400=1かつ昭和59年より前受診、大阪能勢はv1400=1かつ昭和56年度より前受診、大阪熊取はv1400=1かつ昭和58年より前受診。<br>注: ベースライン調査は、宮城は昭和59年1～2月、愛知名古屋は昭和60年10～11月、愛知犬山は昭和60年7～8月、大阪能勢は昭和58年2月、大阪熊取は昭和60年2～3月、大阪河南は昭和59年11～12月、大阪東成は昭和59年10～11月に実施。 |
| v1404 | 数値 | 胃レントゲン検査受診場所(住民健診) | 一部相違 | 1: Yes (欠損=.) | 大阪熊取は「集団検診で(例えば成人病予防協会検診車)」。愛知名古屋は「市医師会健診センターまたは休日急病診療所(検診車)」。 |
| v1405 | 数値 | 胃レントゲン検査受診場所(保健所・保健センター) | 一部相違 | 1: Yes (欠損=.) | 愛知名古屋と大阪東成のみ。<br>v1400=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1406 | 数値 | 胃レントゲン検査受診場所(職場) | 一部相違 | 1: Yes (欠損=.) | v1400=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1407 | 数値 | 胃レントゲン検査受診場所(病院など) | 一部相違 | 1: Yes (欠損=.) | v1400=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1408 | 数値 | 胃レントゲン検査受診場所(農協) | 一部相違 | 1: Yes (欠損=.) | 大阪熊取と大阪河南のみ。<br>v1400=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1409 | 数値 | 胃レントゲン検査受診場所(その他) | 一部相違 | 1: Yes (欠損=.) | v1400=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1410 | 数値 | 胃レントゲン検査受診回数(5年間) | 共通 | (欠損=.) | 大阪東成、大阪能勢、大阪河南は欠損。 |
| v1500 | 数値 | 健康診査または血圧検査受診有無 | 一部相違 | 1: Yes, 0: No (欠損=.) | 愛知は欠損。<br>宮城、大阪能勢、大阪熊取は「血圧の検査」、大阪東成、大阪河南は「健康診査(血圧測定や尿検査)」。<br>宮城は「昨年(昭和58年)1年間に」、大阪東成、大阪河南は「この1年のあいだに」、それ以外は単に「受けたことがありますか」と聞いて下位で時期質問。 |
| v1501 | 数値 | 健康診査または血圧検査受診有無(1年以内) | 一部相違 | 1: Yes (欠損=.) | 愛知は欠損。<br>宮城、大阪東成、大阪河南はv1500=1、大阪能勢はv1500=1かつ昭和57年度受診、大阪熊取はv1500=1かつ昭和59年受診。<br>注: ベースライン調査は、宮城は昭和59年1～2月、愛知名古屋は昭和60年10～11月、愛知犬山は昭和60年7～8月、大阪能勢は昭和58年2月、大阪熊取は昭和60年2～3月、大阪河南は昭和59年11～12月、大阪東成は昭和59年10～11月に実施。 |
| v1502 | 数値 | 健康診査または血圧検査受診有無(1年より前～2年以内) | 一部相違 | 1: Yes (欠損=.) | 宮城、愛知、大阪東成、大阪河南は欠損。<br>大阪能勢はv1500=1かつ昭和56年度受診、大阪熊取はv1500=1かつ昭和58年受診。<br>注: ベースライン調査は、宮城は昭和59年1～2月、愛知名古屋は昭和60年10～11月、愛知犬山は昭和60年7～8月、大阪能勢は昭和58年2月、大阪熊取は昭和60年2～3月、大阪河南は昭和59年11～12月、大阪東成は昭和59年10～11月に実施。 |
| v1503 | 数値 | 健康診査または血圧検査受診有無(2年より前) | 一部相違 | 1: Yes (欠損=.) | 宮城、愛知、大阪東成、大阪河南は欠損。<br>大阪能勢はv1500=1かつ昭和56年度より前受診、大阪熊取はv1500=1かつ昭和58年より前受診。<br>注: ベースライン調査は、宮城は昭和59年1～2月、愛知名古屋は昭和60年10～11月、愛知犬山は昭和60年7～8月、大阪能勢は昭和58年2月、大阪熊取は昭和60年2～3月、大阪河南は昭和59年11～12月、大阪東成は昭和59年10～11月に実施。 |
| v1504 | 数値 | 健康診査または血圧検査受診場所(住民健診) | 一部相違 | 1: Yes (欠損=.) | 大阪東成は「地区健康相談」。<br>v1500=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1505 | 数値 | 健康診査または血圧検査受診場所(保健所・保健センター) | 一部相違 | 1: Yes (欠損=.) | 大阪東成は「保健所や小学校の市民検診」。<br>v1500=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1506 | 数値 | 健康診査または血圧検査受診場所(職場) | 一部相違 | 1: Yes (欠損=.) | v1500=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1507 | 数値 | 健康診査または血圧検査受診場所(病院など) | 一部相違 | 1: Yes (欠損=.) | v1500=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1508 | 数値 | 健康診査または血圧検査受診場所(農協) | 一部相違 | 1: Yes (欠損=.) | 大阪熊取と大阪河南のみ。<br>v1500=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1509 | 数値 | 健康診査または血圧検査受診場所(その他) | 一部相違 | 1: Yes (欠損=.) | v1500=1かつ当該場所の受診有無・受診年いずれか回答あり |
| v1600 | 数値 | 子宮がん検診受診有無 | 一部相違 | 1: Yes, 0: No (欠損=.) | 女性のみ定義。宮城は「昨年(昭和58年)1年間に」、愛知は「昨年(昭和59年)1年間に」、大阪東成、大阪河南は「この1年のあいだに」、それ以外は単に「受けたことがありますか」と聞いて下位で時期質問。 |
| v1601 | 数値 | 子宮がん検診受診有無(昨年) | 一部相違 | 1: Yes (欠損=.) | 女性のみ定義。宮城、愛知、大阪東成、大阪河南はv1600=1。大阪能勢はv1600=1かつ昭和57年度受診、大阪熊取はv1600=1かつ昭和59年受診。<br>注: ベースライン調査は、宮城は昭和59年1～2月、愛知名古屋は昭和60年10～11月、愛知犬山は昭和60年7～8月、大阪能勢は昭和58年2月、大阪熊取は昭和60年2～3月、大阪河南は昭和59年11～12月、大阪東成は昭和59年10～11月に実施。 |
| v1602 | 数値 | 子宮がん検診受診有無(一昨年) | 一部相違 | 1: Yes (欠損=.) | 女性のみ定義。宮城、愛知、大阪東成、大阪河南は欠損。<br>大阪能勢はv1600=1かつ昭和56年度受診、大阪熊取はv1600=1かつ昭和58年受診。<br>注: ベースライン調査は、宮城は昭和59年1～2月、愛知名古屋は昭和60年10～11月、愛知犬山は昭和60年7～8月、大阪能勢は昭和58年2月、大阪熊取は昭和60年2～3月、大阪河南は昭和59年11～12月、大阪東成は昭和59年10～11月に実施。 |
| v1603 | 数値 | 子宮がん検診受診有無(一昨年より前) | 一部相違 | 1: Yes (欠損=.) | 女性のみ定義。宮城、愛知、大阪東成、大阪河南はv1600と同義。大阪能勢はv1600=1かつ昭和56年度より前受診、大阪熊取はv1600=1かつ昭和58年より前受診。<br>注: ベースライン調査は、宮城は昭和59年1～2月、愛知名古屋は昭和60年10～11月、愛知犬山は昭和60年7～8月、大阪能勢は昭和58年2月、大阪熊取は昭和60年2～3月、大阪河南は昭和59年11～12月、大阪東成は昭和59年10～11月に実施。 |
| v1604 | 数値 | 子宮がん検診受診場所(検診車・住民検診) | 一部相違 | 1: Yes (欠損=.) | 女性のみ定義。大阪東成、大阪能勢、大阪熊取は欠損。大阪河南は「河南町が行った集団検診」。宮城、愛知は「検診車」。<br>v1600=1かつ当該場所の受診有無・受診年いずれか回答あり |