平成 26 年度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：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、データを広く共有化し二次利用を促進するためのシステムであるデータアーカイブ化が進められている社会科学系ならびにライフサイエンス系のデータアーカイブセンターの現状を把握した。疫学研究データには機微情報を含むのみならず、収集する項目数も多いことから、80%程度のレコードは一意性があるものとして対応することが必要であり、研究開始時点での対象者への説明のあり方、完全に連結不可能匿名化にするタイミングやその方法などの検討とともに、研究内容によっては共同研究を締結して行うなどの対応が必要と考えられた。また、今後、疫学研究のデータアーカイブを構築し利用を進めていくためには、教育も含めた利用環境整備も必要である。一方、統計法の規定から、現状では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利用することはできない。そのため、がん登録推進法による死亡者情報票の活用、ならびに現制度化で運用するために必要に応じて死因情報を入手・利用する方法を提案した。

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。今年度は、①国内における社会学分野ならびにライフサイエンス分野のデータアーカイブの現状、②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検討した。さらに、人口動態統計情報を利用して把握されている死因情報は、統計法の規定により、データアーカイブ時には制限つきであれ公開情報とできないことから、その対応として代替え案として、③がん登録推進法において示された死亡者情報票の活用の提案、ならびに④現制度化での利用方法の提案を行った。

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アーカイブを公開する場合の課題を検討した。

②三府県コホートデータを用い、分類数、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。
③がん登録推進法について、条文や政令をもとに、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、死亡者情報活用をめぐる諸問題について検討を行った。
④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。

C. 研究結果

①-1 社会学分野のデータアーカイブの現状

データアーカイブセンターの意義は、統計調査、社会調査の個票データを収集・保管し、その散逸を防ぐとともに、学術目的での二次的な利用のために提供することにある。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。データアーカイブを二次利用するメリットは、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、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、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。

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。

① -2 ライフサイエンス分野のデータアーカイブの現状

バイオサイエンスデータベースセンター（NBDC）では特にヒトに関する情報に特化した NBDC ヒトデータベースを構築し、2013 年 10 月から運用が開始された。ヒトを対象とするデータであるため、特に個人につながる情報の保護対策が重要となる。そこで NBDC では、欧米のデータベースを参考に受け皿づくりが進められている。取り扱われるデータは匿名化されたもののみで、レベルに応じたアクセス制限が行われている。データ提供と利用に関する審査は、NBDC で行われるが、原則として試料提供者からデータ共有に関する同意を事前に受けておくことが求められている。ただし、過去に収集された既存試料・情報で同意を取り直すことが困難な場合には、データ共有について倫理委員会で承認されていることが要件である。データ共有に関するガイドライン、セキュリティレベルに関するガイドライン等が定められ、HP 上で公開されている。2015 年 2 月現在、研究データ 15 件が HP に公開され、そのうち制限公開 10 件、オープン 5 件であるが、すべてゲノムに関連するもので、いわゆる疫学研究のデータは今までのところ、寄託はされていない。

② 一意性の検討

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 がん登録推進法における死亡者情報票の活用

死亡者情報票を利用したコホートデータ追跡情報入手法を検討するため、癌登録推進法における死亡者情報票の取り扱いについて確認した。がん登録推進法第 11 条（死亡者情報票の作成及び提出）に死亡者情報票の作成に関する事項は明記されている。それによると、死亡者情報票は、死亡の届書その他の関係書類に基づいて、市町村長が作成するもので、死亡した者に関する氏名、性別、生年月日、死亡の時における住所、死亡の日、死亡の原因、死亡診断書の作成に係る病院又は診療所の名称及び所在地その他の厚生労働省令で定める情報が含まれる。全死亡者に関する死亡者情報票は、電磁的記録又は書類により作成され、保健所、都道府県を経て、国（国立がん研究センター）に提出される。それを受けて、国立がん研究センターは、死亡者情報票と全国がん登録情報とを照合する。その照合期間は、厚生科学審議会がん登録部会の政令案では 100 年とされたことから、がん患者の生命予後をほぼ完璧に追跡することが可能になったと思われる。

④データアーカイブ化における死因情報の利用：コホートデータに死因情報を付与する方法

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利用することは不可能な状態にある。そこで、代替え案の一つとして、死因を連結した形でのデータセット構築・アーカイブ化ではなく、必要時に中央（アーカイブデータを保持するセンター等）で死因照合作業を行って解析用データセットを作成する方法とその妥当性を JALS の実データを用いて検討した。

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。

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。実際、社会科学系分野では、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。ライフサイエンス分野では、NBDCがデータアーカイブセンターの役割を担い始めた。しかし、いまだ緒についたところで、特にヒトを対象としたデータに関して実績があがるのはまだこれからと考えられた。加えて。疫学研究は単に生体情報のみならず、生活習慣や心情等に関する情報も収集されることが多く、追跡結果も死亡・疾病罹患など機微情報を含む。さらに、収集する項目数も多いため、その組み合わせにより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。したがって、研究開始時点での対象者への説明のあり方、完全に連結不可能匿名化にするタイミングやその方法など、今後の検討課題である。また、単に塩基配列などの公開と異なり、疫学研究で構築されたデータセットの公開内容は一部に制限し、機微情報を取り扱う場合には共同研究を締結するなどの対応が必要と思われる。

統計法の規定から、現状では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利用することはできない。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。がん登録推進法により提供される死亡者情報票データを活用して、米国のNDIと同様のシステムを作るには法制度の改革が必要であるが、それが実現すれば、わが国の疫学研究・臨床研究や医薬品・医療機器開発は発展すると思われ、それは政府「健康・医療戦略」の目指すところと合致するものであろう。

一方で、現制度化での運用方法を検討するため、死因情報を外したアーカイブ環境を想定し、必要時に死因を人口動態二次利用申請し、アーカイブセンターにて照合・集計・解析を行う運用例を提案した。JALSで実際に死因照合を行ったところ99.5%で照合が可能であり、照合作業の技術的側面、作業手順化の面で問題はなかった。今後検討すべき課題の一つとしては、人口動態統計二次利用申請に基づくことから、データの保持期間が公的研究費の継続期間に限定されることがあげられる。研究の質や結果の再現性を保証するという点では、解析に使用した死因つきのデータセットが長期に保持できることが望ましい。また、今回提案する方法では、死因つきデータセットの利用場所、すなわち解析場所がアーカイブセンター（あるいは申請書に記載した研究者の所属する機関）に限られる。そのため、アーカイブデータの利用規定も合わせて、アーカイブセンターで対応する場合は、データ解析を行える環境（物理的な環境、統計家の配置等）について検討する必要がある。また、死因データの申請者の所属機関で実施する場合には、アーカイブデータの外部利用の規約等の整備も必要があるといえる。

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課題を整理するために、社会科学系ならびにライフサイエンス系のデータアーカイブセンターの現状を把握した。疫学研究データには機微情報を含むのみならず、収集する項目数も多いことから、80%程度のレコードは一意性があるものとして対応することが必要である。研究開始時点での対象者への説明のあり方、完全に連結不可能匿名化にするタイミングやその方法などの検討とともに、共同研究を締結するなどの対応が必要と思われる。また、疫学研究のデータアーカイブを構築していくためには、先行する社会科学系データアーカイブの運営システムから学ぶと同時に、利用のための環境整備も必要と考えられた。一方、統計法の規定から、現状では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利用することはできない。そのため、がん登録推進法による死亡者情報票の活用、ならびに現制度化で運用するために必要に応じて死因情報を入手・利用する方法を提案した。

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