

## 総説

## 公表された既存データの有効活用について

大木秀一<sup>1 §</sup>, 彦 聖美<sup>1</sup>

## 概要

公表された既存データの有効活用について検討した。ここでは容易に入手可能なデータを想定している。この種の研究には、疫学における記述疫学研究および生態学的研究などがある。既存の統計データによる研究の課題点は、これまでは学術的（理論的）な側面を中心に検討されてきた。しかし、保健医療系の学部における研究など、教育的な側面を考えると、時間と経費の節約など多くの利点があると思われる。保健医療系の研究において既存データを活用した仮説設定が、その後の大きな研究の端緒になった例を示した。また、健康の都道府県格差の是正を考える際に有効活用できることなど今後の活用例を提示した。研究目的に応じて有効に活用できるデータは、統計データに限らず多数公表されている。自験例を基に、公表された既存データの活用例を具体的に示した。基本的な考え方自体は、研究テーマが異なっても共通する部分が多いと考える。

キーワード 公表された既存データ, 有効活用, 生態学的研究, 仮説設定, 保健医療系の学部教育

## 1. はじめに

調査研究をするにあたっては、新たにデータを入手しないとイケない、あるいは新たに入手したデータを分析するのが本来の研究であるという考え方が一般的には強いように思われる。これは、専門的な研究だけでなく、教育的な研究（卒業研究や修士課程の研究）についても同様である。確かに、初学者にとって、実際に人からデータを取り、分析するというプロセスを経験することは有意義である。しかし、卒業研究のように限られた時間と予算を考えた場合に、新たにデータを入手するだけではなく、公表された既存データを有効に活用するという選択肢もあると思われる。

保健医療系で、公表された既存データ（主として統計データ）を活用する方法論を詳しく解説した書籍は少ない。未だ研究としての系統的な方法論が確立していないだけでなく、研究方法論としてもあまり認知されていないように思われる。

公衆衛生学の一領域である疫学では、既存の統計データを分析する記述疫学研究<sup>1)</sup>や生態学的研究<sup>1-9)</sup>が一つの方法論として確立している。しかし、広く保健医療分野全般の研究について、既存データを分析する意義について論じられることは少ないように思われる。そこで今回、保健医療系の教育的な側面を含めて検討した。

## 2. 公表された既存データについて

## 2.1 本論における議論の対象

ここでは、インターネット上あるいは文献上に公表され、比較的容易に入手可能であり、利用にあたっては、出典を明記すれば、通常はそれ以上の特別な申請を必要としないデータを想定している。従って、既存でも未公表の健診データやカルテ情報などは該当しない。公表された既存データは統計データであることが多いが、文章データでもよい。本論では簡便で有効に活用できる可能性があるという観点から考えている。

公表された既存データの分類例を表1に示す。統計データの多くは集団レベルの集計データ（集合データ aggregate data）であるが、個人レベルの数値データでもよい。代表的なものは、e-Stat（政府統計の総合窓口）上に公開されたデータ、あるいは学会や民間企業が公表するデータである。ただし、活用にあたっては、調査・集計方法の詳細がある程度公表されている必要がある。既存統計資料と呼ばれることもある<sup>10)</sup>。

なお、既存データの活用の中には、政府の指定統計調査や国民健康保険のレセプト情報の目的外利用、あるいは社会調査データのアーカイブの利用など、一般に二次利用ないし二次分析と呼ばれるものがある。これらはデータの利用にあたって何らかの手続きや制限、場合によっては経費が伴う。また、主たる目的が個人レベルのデータを入

<sup>1</sup> 石川県立看護大学 § コレスポンディングオーサー

表1 公表された既存データの分類例

データの分類		例	
量的データ	集合データ	e-Stat (政府統計の総合窓口) 報告書・年報・白書など	人口動態統計, 患者調査, 国民生活基礎調査 厚生労働科学研究報告書, 体力・運動能力調査報告書, 厚生労働白書, 文部科学白書, 環境白書, 感染症サーベイランス, がん登録
	個人単位データ	学会誌に掲載されるデータなど 都道府県別各種年報など 民間企業の調査報告など 学会の委員会報告など	厚生の指標, 日本産科婦人科学会雑誌 県統計書, 県統計年鑑 8020推進財団, ベネッセ教育総合研究所 生殖補助医療実施後の先天異常の発生事例 (日本産科婦人科学会)
質的データ		新聞記事 住所録・名簿など 文献(論文)タイトル一覧	虐待(死), 事故・自殺などに関する報道記事 各種保健医療機関名簿 学会誌索引号 (ないし学会サイト内の情報)

\* 学術論文(一般に文献)の内容は量的データとも質的データともなりうる。

手して再分析することなので, 今回の議論の対象には含めない。

統計データ以外にも, データベース化されることは少ないが, 研究者のアイディア次第で有効活用できる公表されたデータが多数ある。そのすべてを系統的に述べることは難しいが, 例えば, 未集計の健康・疾患情報のリストや住所録に類するもの, 新聞記事などがあげられる。この中で新聞記事の分析は保健医療分野でも比較的多い。また, 学術論文に含まれる情報も公表されたデータとみなすことができる。文献レビュー論文は論文データを有効活用した研究の一類型といえる。

## 2.2 特徴別に見た統計データの種類

研究者の関心あるテーマや変数(要因)によっては, 集団レベルのデータしか得られないことも多い。集計データは3種類に大別できる<sup>2,7,9)</sup>。

(1) 集団の統計量(個人データから算出したもの): コンテキスト(文脈)変数ともいわれる。例えば, 集団における, 平均食塩摂取量や喫煙率である。これを個人データの代用とする。個人データの統計量は, 集団レベルになると, 様々な社会・自然環境要因の影響を受けて新たな意味を持つ。従って, データの持つ文脈の理解が必要になる。

(2) 個人データが入手・測定困難: 環境測定値(例えば, 大気汚染曝露量, 日照時間)で, 個人レベルでの測定が困難な場合には, 個人レベルでの測定評価を模索しながら集団データで代用する。長期縦断データの代用として, 人口動態統計で出生コホートを仮定するのも同様の考えによる。

(3) 集合データ以外には存在しない: 人口密度や地域面積, あるいは健康に関する法の有無や健康政策などが該当する。

## 3. 公表された統計データの分析

### 3.1 統計データのデータベース化

公表される統計データは電子媒体のこともあれば紙媒体のこともある。いずれにしても, これを新たに利用する研究者の目的に沿って提供されているわけではない。従って, 分析にあたっては目的に合わせてExcelなどでデータシートを作成する必要がある。

毎年公表されるデータは, 一度フォーマットを決めてExcelに入力し, 集計をExcel関数で行っておけば, データを追加することやその後の集計が容易であり, 経年変化を観察しやすい。

### 3.2 分析の方法

研究計画を考えることは他の研究デザインと同じであるが, データの入手可能性を考えながら計画を修正したり, データに合わせた分析方法を検討する必要がある。無計画に統計データを分析しても, 他の研究結果との比較可能性が維持できず, 無駄になる可能性がある。ここでは, 基本的な疫学的分析方法<sup>1,8,9)</sup>に沿って説明する。

#### (1) 記述疫学研究

記述疫学では, 明確に定義された集団における健康事象の分布と頻度を「人」, 「場所」, 「時間」の3要素について記述する。記述疫学は原則的には全数調査か大規模な標本抽出調査であるから, かなりの時間と労力を要する。その意味で, 大半は個人の努力で実施できるものではない。

「人」に関する要因として必須なものは, 性と年齢である。海外ではこれに人種が加わる。

「場所」に関する要因としては発生場所, 例えば各国別, 都道府県別の頻度差(地域差)を記述する。疾病・健康事象の地理的分布の記述は, これに関係する要因についての仮説設定や実際の予

防対策を行う上で重要である。地図上に表すと非常にわかりやすくなる（疾病地図）。これにより、疾病の多発地域や地域集積性が見出されることもある。近年のGIS（Geographic Information Systems）の発展に伴い、地理情報に特化した専門分野に空間疫学<sup>11)</sup>がある。

「時間」に関しては短期傾向から長期傾向まで様々である。年次推移、周期変動、季節変動、月別変動などに注目する必要がある。数十年単位の観察により将来予測が可能な場合もある。

時間の分析では、社会学から発展してきたコホート分析（APC分析：Age-Period-Cohort Analysis）<sup>112)</sup>がある。これは、長期的なデータから、その変化の要因を加齢の要因による影響（年齢効果 age effect）、時代の要因による影響（時代効果 period effect）および世代差の要因による影響（コホート効果・世代効果 cohort effect）に分離するものである。何歳か（年齢効果）だけではなく、いつ（時代効果）、どのような時代に（世代効果）生まれたかが健康事象には影響する。これは、戦争や自然災害、飢饉などの身体的・精神的な影響を集団レベルで観察する場合にも用いられる。

記述疫学研究では性や年齢など一部の変数を除いて、健康事象に関係する要因（原因）の情報が得られないという制約がある。しかし、厚生労働省の人口動態統計など既存の優れた資料を利用すれば比較的容易に国レベルの記述が可能である。

#### （2）生態学的研究

生態学的研究とは、既存資料を利用して集団レベルでの健康事象（結果）とその要因（原因）の関心の地理的分布、時間的分布を比較する研究方法である<sup>1-9)</sup>。

生態学という言葉が使用されるのは、この研究方法では人間とそれを取り巻く環境についての総合的な情報を収集し観察するからである。個人ではなく集団が基本単位だということが最大の特徴である。集団レベルでの社会的要因（文化、経済、風習、教育など）、地理的要因（高度、緯度、気候など）の分布と健康事象の関係を調べる場合に有用である。主要な解析方法は2つある。

第1は、異なる国や地域（例えば、都道府県）において、ある一定の時期の健康事象とそれに関わる要因との関係を比較検討する方法であり、地域相関研究と呼ばれる。統計学的には散布図を描き、必要に応じて相関係数、回帰直線を算出する。

第2は、特定の国や地域において、健康事象と

その要因の時間的変化を同時に観察し、両者の関係を検討する方法であり、時系列研究と呼ばれる。地域と時間を複合した分析も可能である（複合研究）。以上を基本に、より高度な分析手法が多数開発されているが、ここでは割愛する。

集団レベルの分析結果で問題になるのが、生態学的錯誤（生態学的誤謬）<sup>8,9)</sup>である。これは、集団で得られた結果が必ずしも、個人レベルでは当てはまらない、時には正反対になることもあるというものである。集団レベルの分析結果の解釈については、今でも理論面で多くの議論がある<sup>1-9)</sup>。集団レベルでの議論か、個人レベルでの議論かを意識的に区別して考えることが重要である。

### 3.3 既存の統計データ分析の利点と課題点

公表された統計データを研究目的で活用することに関しては、保健医療系の分野ではこれまで学術的（理論的）側面からの議論が中心であり<sup>1-9)</sup>、教育的側面からの議論は必ずしも多くない。そのような中で、社会学の分野では教育的側面からの議論や活用がいくつかなされている<sup>13,14)</sup>。

既存の統計データを活用する際の利点と課題点を学術的・教育的側面に焦点を当てて表2にまとめた。統計データの分析における課題点の多くは、これを個人レベルの要因分析の代用に用いることに付随する理論上の問題である。研究目的を集団レベルでの記述的内容（記述疫学）、あるいは集団レベルでの効果測定などに設定すれば問題は少ない。また、保健医療系の学部教育的な側面においても多くの利点があるといえる。

### 4. 公表された統計データを分析する意義

これは統計データを利用する利点とも重複する。統計データの分析はバイアスの問題を含め、要因分析において個人レベルの因果関係の決め手になることは少ない。しかし、仮説設定という意義を考えた場合には、重要な役割を果たし得る。統計データの分析が端緒となり、その後研究が大きく推進した事例と今後の活用において有効と思われる分野を簡単に紹介する。

#### 4.1 先駆的事例

##### （1）成人病胎児期発症起源説

胎内での低栄養状態が成人期以降の生活習慣病などの発生に影響するという仮説である<sup>15)</sup>。21世紀最大の医学仮説ともいわれるこの仮説の提唱

表2 既存の統計データを活用する際の利点と課題点

側面	学術的側面でも教育的側面でも共通する点	保健医療系の教育的側面での特徴
研究方法	<ul style="list-style-type: none"> <li>・時間制約が少なく、研究者主導で研究できる</li> <li>・比較的短時間で結果を出せる</li> <li>・全国的規模の優れたデータを有効活用できる</li> <li>・データを入手困難・調査が困難な研究テーマでも対応できる</li> <li>・データの質が研究者に依存しない</li> <li>・集合データを用いる以外には困難な研究を実施できる</li> </ul>	<ul style="list-style-type: none"> <li>・生態学的に異なるレベル（集団と個人）での現象の違いに対する理解を深めることができる</li> <li>・保健師国家試験の出題基準と合致する（保健統計調査、疫学調査法、地域診断の過程など）</li> <li>・多くの場合複雑な分析を必要としない</li> <li>・データ分析や論文執筆の教育に（相対的に）時間を多く配分できる</li> </ul>
分析	<ul style="list-style-type: none"> <li>・個人では調査不可能な記述疫学統計を提供でき、全国データであれば選択バイアスが殆どない</li> </ul>	<ul style="list-style-type: none"> <li>・ある程度確実な結果が期待できる</li> </ul>
結果公表	<ul style="list-style-type: none"> <li>・政策やプログラムの介入効果を集団レベルで評価するときに有用である</li> <li>・仮説形成に有用であり、有意な結果がさらなる研究につながる可能性がある</li> </ul>	
研究倫理	<ul style="list-style-type: none"> <li>・問題が少ない</li> </ul>	<ul style="list-style-type: none"> <li>・データ収集を通じて情報リテラシーを磨ける</li> </ul>
研究モラル	<ul style="list-style-type: none"> <li>・学術論文として研究結果を公表できなくても（人を対象としてデータを収集した研究に比べて）研究モラルに反することはない</li> </ul>	
課題点	<ul style="list-style-type: none"> <li>・健康事象とその要因の因果関係を証明する力は弱い</li> <li>・健康事象とその要因が異なる集団からのデータとなりやすい</li> <li>・集団で当てはまる結果が個人レベルで当てはまるとは限らない（生態学的誤謬）</li> <li>・交絡因子を制御できないことが多い</li> <li>・オリジナリティが低いとみなされやすく、研究として低く評価される可能性がある</li> <li>・データそのものの妥当性の問題がある</li> <li>・検出力が落ちやすい</li> <li>・集団の人口規模の違いの影響がある</li> </ul>	<ul style="list-style-type: none"> <li>・人を対象とした研究プロセスを習得する機会を経験できない（例えば、倫理申請、質問紙の作成と発送など）</li> <li>・データの背景にある現実感覚が希薄になる可能性がある</li> <li>・データ収集を軽視する可能性がある</li> </ul>

文献1-10,13)などを参考に作成した。

者である Barker によれば、人生初期の発達が成人期の心血管疾患発症に重要である可能性を解く鍵は、意外にも地域相関研究にあった<sup>16)</sup>という。即ち、イングランドとウェールズにおける多数の地区を対象に、1世紀前の新生児死亡率と今日の心疾患死亡率の散布図を作成した結果、両者に強い相関を見出したのである。この関係は疾病地図からも確認できる。この仮説はその後 DOHaD (Developmental Origins of Health and Disease) 仮説として大きく発展する。対象疾患も大幅に拡大し、動物実験も含めて分子レベルの研究から集団レベルの研究まで、膨大な数の研究を引き起こし今日に至る。

#### (2) フレンチパラドックス

フランスでは飽和脂肪酸が豊富に含まれる食事を摂取しているにもかかわらず、冠動脈疾患死亡率が低いことをフレンチパラドックスという。この用語自体は1992年に Renaud ら<sup>17)</sup>が最初に用いているが、生態学的研究による類似の研究報告はそれまでも欧州で相次いでいた。フレンチパラドックス発見のきっかけは、各国の食事の脂肪摂取量と冠動脈疾患死亡率の地域相関研究である。なお、赤ワイン（一般には適度の飲酒）を飲むことで心疾患発症のリスクが低減するというの

は一つの解釈であり、フレンチパラドックスそのものではない。こうした、地域相関研究が端緒になり冠動脈疾患と生活習慣（食習慣、運動習慣など）との関連についての様々なレベルでの研究がその後大きく進展した。

#### 4.2 今後の活用の例

21世紀の国民健康づくり政策である健康日本21は、2013年4月より第二次を迎えた。その中では「第一 国民の健康の増進の推進に関する基本的な方向、一 健康寿命の延伸と健康格差の縮小」を謳っている<sup>18)</sup>。そこでは、「あらゆる世代の健やかな暮らしを支える良好な社会環境を構築することにより、健康格差の縮小を実現する」<sup>18)</sup>としている。ここで健康格差とは、「地域や社会経済状況の違いによる集団間の健康状態の差」であると定義している。健康格差の縮小とはより具体的には、健康寿命（健康上の問題で日常生活が制限されることなく生活できる期間）の「都道府県格差」を縮小することである。即ち、個人レベルでの疾患予防だけでなく、集団（都道府県）レベルでのプログラム（介入）効果の評価が必須である。

こうした考え方の起源は、Rose<sup>19)</sup>の先駆的な

論文に見ることが出来る。個人レベルでの健康決定要因だけでなく、集団レベルでの健康（罹患率低下）決定要因があり、それぞれがハイリスクアプローチとポピュレーションアプローチに対応する。両者は必ずしも対立するものではないが、まず罹患率低下を考えるべきであるという主張である。つまり、ある集団における個人差の決定要因と集団間の差の決定要因は異なる可能性があり<sup>20)</sup>、現在のような個人レベルの研究を偏重すると、罹患率低下（一次予防）を果たせない可能性がある<sup>6)</sup>。

健康の社会的決定要因<sup>21)</sup>の多くは個人レベルの変数であると同時に集団レベルの変数でもある（例えば、交通手段・社会的排除・食糧・ソーシャルサポート）。個人を対象とした健康教育を熱心に行っても、生活習慣病予防が必ずしも奏効していない現状を考えれば、集団レベルの研究や、それを基にした健康戦略がより注目されてもよいだろう。既存の統計データの有効活用は、今後この種の研究に寄与すると考えられる。

#### 4.3 その他の活用例

これ以外にも、人口動態統計、患者調査、国民生活基礎調査など全国データあるいは都道府県別データを分析することで、健康施策の優先順位の決定、地域診断、健康施策やプログラム（介入）効果の測定、健康事象の要因解明、保健医療資源の将来需要予測など様々な研究が可能である<sup>10)</sup>。

#### 5. 公表されたデータの具体的な分析例

以下に、自験例<sup>22,23)</sup>を、公表されたデータの有効活用という視点から具体的に示す。研究テーマは、多胎出生の動向と背景因子、多胎出生の公衆衛生学的な影響などである。今回の執筆にあたり、最新のデータを用いて再分析した。紙面の都合で文献の引用は最小限にとどめた。解説がやや長くなるが既存データを用いた系統的な分析の例と考えていただきたい。研究テーマが変わっても、基本的な考え方は共通していると考えられる。

主に用いたデータソースは、①人口動態統計（厚生労働省：1951年～2012年分）、②「倫理委員会登録・調査小委員会報告（体外受精・胚移植等の臨床実施成績および登録施設名）」（以下、生殖補助医療に関する報告書）（日本産科婦人科学会：1985年～2011年分）、③「子ども虐待による死亡事例等の検証結果等について」（以下、虐待死亡事例の報告書）（厚生労働省：第1次～第

9次：2003年7月～2012年3月）、④インターネット上の新聞記事、である。①は年ごとに冊子体、CD-ROMないし一部は電子データとしてインターネット上（e-Stat）に提供されている。②と③は、インターネット上ではpdfファイルとして提供されている。②は調査対象医療施設の回収率が1999年以降ほぼ98%以上であり、全国悉皆データとみなせる。④は有料のデータベースも存在する。

なお、公表されたデータの特徴に合わせて研究テーマを選択することと、研究テーマに合わせて新たなデータソースを探し出すことは、双方向に進むものである。実際にはこれ以外のデータソースも適宜使用している。統計データの分析にあたっては、必要な年次分のデータを収集した上で、Excelのデータシートに分析に適した形式で再入力している。

#### 5.1 集団レベルのデータの活用例

##### (1) 多胎出生の妊娠の方法別年次推移<sup>22,23)</sup>

出生に関する最大の公表データは人口動態統計である。過去62年分の人口動態統計を基にまず、多胎の出生割合（出生1,000に対する多胎児の数）の年次推移を調べた（図1）。出生割合は1951

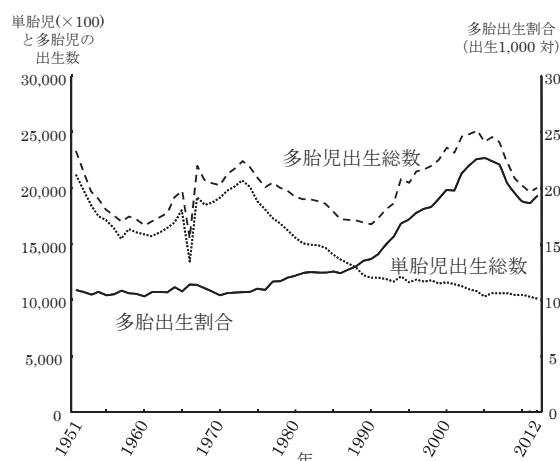


図1 多胎児と単胎児の出生数および多胎出生割合の年次推移（人口動態統計を基に作成）

年～1976年では11前後で推移するが、1980年代後半以降急増し2005年にはピークに達し23と倍増し、その後、減少に転じる。

まず、2005年までの多胎出生割合の倍増について検討した。不妊治療において多胎妊娠が増加することは当然予想される。しかし、人口動態統計には妊娠の方法（自然妊娠か不妊治療妊娠か）

に関する情報はない。そこで、自然の多胎出生頻度が母親の年齢と共に増加するという生物学的事実を根拠に以下の推定を行った。年ごとの自然の多胎出生数を推定するために、不妊治療が普及する前である1974-1976年全体での母親の年齢階級別多胎出生割合を基準値として算出し、その後の年次の年齢階級別総出生数にこの基準出生割合を乗じて、その年次の年齢階級別自然多胎出生数を推定した。これを合計すれば年次ごとの自然多胎妊娠出生数が推定でき、さらに多胎出生全体から自然多胎妊娠出生数を引けば、不妊治療妊娠による多胎出生数が推定できる。以上の出生数を出生割合に換算する。この推定方法は自然の多胎出生数をやや多めに見積もるが、その影響は小さい。

分析の結果(図2)を見ると、過去35年の多胎出生割合の増減は不妊治療による多胎出生割合の増減とはほぼ一致する。自然の多胎出生割合は殆ど増減がない。従って、多胎出生が増加した主たる原因は不妊治療だと推定される。

(2) 不妊治療の種類別多胎出生<sup>23)</sup>

不妊治療による妊娠は、体外受精・胚移植、顕微授精などの生殖補助医療(ART: Assisted Reproductive Technology)とそれ以外の一般不妊治療(主として排卵誘発)に大別される。ARTについては、日本産科婦人科学会が施設登録制をとっており、毎年行う実態調査の集計結果の一部(ARTの種類、妊娠数、分娩数、出生数など)を学会誌で報告論文(生殖補助医療に関する報告書)として公表している。

一般不妊治療はART登録施設以外でも広く

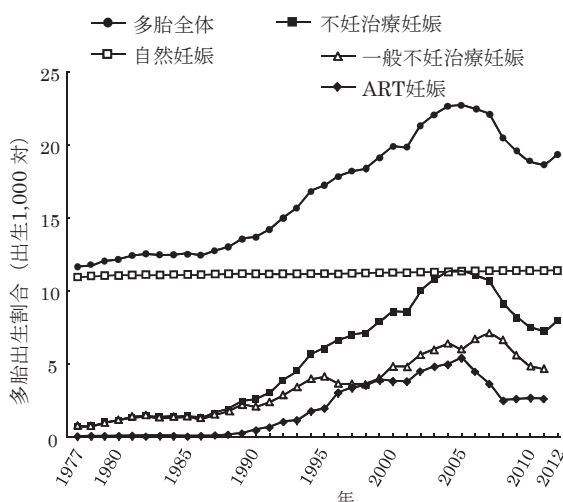


図2 妊娠の方法別に見た多胎出生割合の年次推移  
ART: Assisted Reproductive Technology  
(人口動態統計, 生殖補助医療に関する報告書を基に作成)

実施されており、その実態は殆ど不明である。ARTによる多胎出生数は2007年から実数が公表されている。それ以前のART多胎出生数は、全ART出生数(児の数)と全ART分娩数(母の数)を基に作成した推定式などで算出した。前述した不妊治療による多胎出生総数からARTによる多胎出生総数を引けば、一般不妊治療による多胎出生総数が推定できる。以上の出生数を出生割合に換算する。不妊治療に限れば、ARTよりも一般不妊治療による多胎出生が多いと推定される(図2)。

(3) 単一胚移植の効果<sup>24)</sup>

周産期のトラブルが母児に多発する多胎妊娠予防に確実な方法の一つは、ARTで胎内に戻す胚の数を制限することである。現在、先進国の多くは単一胚移植(SET: Single Embryo Transfer)を推奨している。日本では、日本生殖医学会が2007年にSETのガイドラインを公表し、これを受けて日本産科婦人科学会が2008年にSETの見解を示している。しかし、ARTによる多胎出生割合は前述の通り2005年をピークに減少しており(図2)、ガイドライン公表の以前からSETの動きは進んでいたと思われる。以上を踏まえ、SETの普及と多胎妊娠に対する影響を分析した。

生殖補助医療に関する報告書では2007年以降SETの成績を報告している。2007年から2011年の値をプロットした結果、ARTにおいてSETによる妊娠の占める割合が増加すると共に、多胎

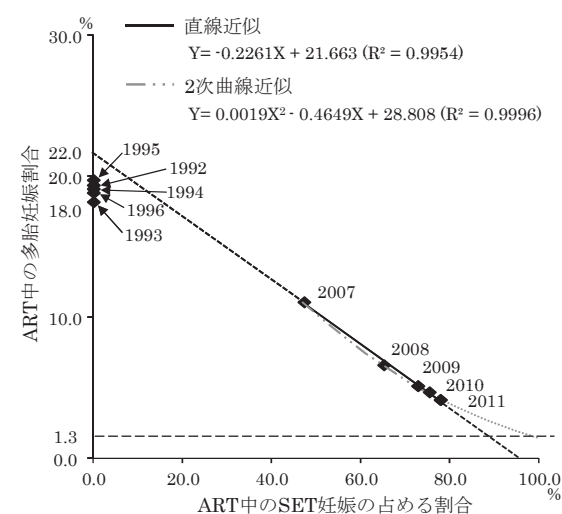


図3 ART中のSET妊娠の占める割合とART中の多胎妊娠割合の変化  
ART: Assisted Reproductive Technology  
SET: Single Embryo Transfer  
(生殖補助医療に関する報告書を基に作成)

妊娠の割合がほぼ直線的に減少している（直線近似で  $R^2=0.9954$ ）（図3）。SETが極めて確実な多胎妊娠予防効果を示したことが確認された。

文献的に見ると1992年から1996年頃はSETが殆ど普及していない。この時期のARTによる多胎妊娠の占める割合は18-20%であり（図3）、複数胚移植での多胎妊娠の割合と考えられる。

実際には、ART妊娠全体に占めるSETによる妊娠の割合が増加すると共に、多胎妊娠の割合は非常に緩やかな下に凸のカーブを描いて減少した（2次曲線近似で  $R^2=0.9996$ ，図3）。仮に、このまま減少すると、ART妊娠のすべてがSETによる妊娠になっても多胎妊娠が1.3%程度残る。一般には、不妊治療で増加するのは二卵性双胎だと考えられているが、不妊治療により一卵性双胎も増加することが度々報告されてきた。最近のシステマティックレビューとメタアナリシスによれば、ARTを受けた妊婦全体のおよそ1%弱が一卵性双胎を妊娠するとされ、自然妊娠の2倍以上に高頻度である。SETに限れば、その頻度はさらに高い(1.5%前後)と推定されている。症例データとは異なる方法で推定した今回の結果も同様の値であった。

(4) 多胎出生の低出生体重に対する影響<sup>25)</sup>

近年の低出生体重児の増加は母子保健領域での最重要課題の一つである。多胎出生が低出生体重に与える影響を経年的に分析した。具体的には、人口動態統計を基に単胎・多胎における低出生体重の割合を求め、これと前述の多胎出生割合から相対危険と集団寄与危険割合を推定した。その結果、出生体重1,500g未満で多胎の影響が最大であることが明らかとなった（図4）。この種の分

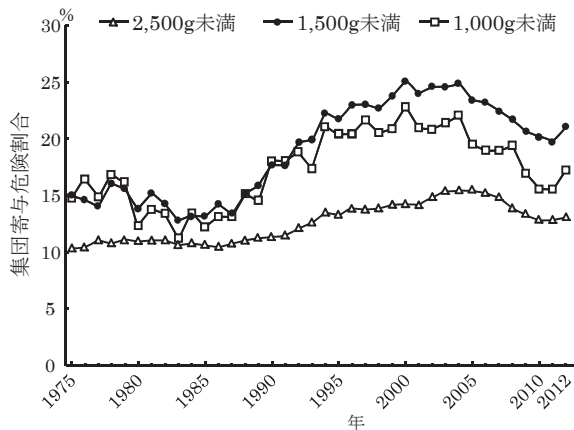


図4 出生体重群別にみた集団寄与危険割合の年次推移 (人口動態統計を基に作成)

析は、早産、死産、乳児死亡など人口動態統計で公表されている他の母子保健指標にも応用できる。

(5) 多胎育児と虐待死亡事例<sup>26)</sup>

多胎児ないし多胎児家庭では虐待が発生しやすいとされるが、全国的なデータは殆どない。そこで、虐待死亡事例に注目して多胎のリスクを検討した。死亡統計は一般に有病・罹患統計よりも入手しやすく、誤差が少ないことを利用した。

虐待死亡事例の報告書では、18歳未満の虐待死亡事例の数および、家庭の要因、児の要因、育児環境などの背景因子が集計結果として掲載されている。全調査期間に495児の虐待死亡事例（心中を除く）がみられ、多胎児は14児存在した。人口動態統計を基に出生コホートを想定し、調査期間に生存した0歳から17歳までのすべての単胎児と多胎児それぞれの生存人年を見積もり、単胎児と多胎児に分けて虐待死亡事例の発生頻度および相対危険を推定した。

虐待死亡事例の報告書では、全般的に欠損値が多い。相対危険を過大評価しないために、欠損値を含めた分析では、欠損値を単胎児とみなした。多胎児は単胎児に比べて、1.6倍（欠損値を除くと2.2倍）虐待死亡の発生頻度が高いと推定された。家庭あたりで見れば多胎児家庭では発生頻度が3~4倍高くなる（児が平均2倍弱多いため）。この値は10代妊娠（人口動態統計の出生時母親年齢20歳未満で代用）の相対危険（13.9と22.4）よりかなり低く、児あたりでは低出生体重児の相対危険（1.4と2.8）とほぼ同じであった。多胎出生と低出生体重は重複するリスク要因だが、集合データでこれを分離することはできない。

5.2 個人レベルのデータの活用例

(1) 単胎児と多胎児での先天異常の比較<sup>27-29)</sup>

日本では、先天異常は新生児死亡・乳児死亡の死因の第一位を占めている。国内には、全国悉皆の先天異常の登録はない。また、人口動態統計には不妊治療の情報がない。そのような中、生殖補助医療に関する報告書で2004年以降、すべての先天異常の症例の疾患名、胎児数（単胎・双胎など）、妊娠期間、母親年齢、児の性別、予後などが一覧表として報告されている点に注目した。

報告書のすべての症例をExcelに入力し、データベースを作成した上で、国際疾病分類（ICD-10, 2003年版）を基にコードQ00-Q99に該当する疾患のみを先天異常として抽出した。

次いで、疾患分類の大分類、中分類、一部疾患の小分類別に単胎妊娠と多胎妊娠における発生頻度と相対危険を推定した。ARTによる単胎妊娠総数と多胎妊娠総数は2006年までは前述した推定値を、2007年以降は報告値を用いた。

先天異常全体に関しては、妊娠あたりでは、多胎妊娠において有意に発生頻度が高かった。出生あたりでは単胎児と多胎児で有意な差は見られなかった。この違いは多胎妊娠では、1妊娠で2児以上になるためである。「中枢神経系の先天異常」は妊娠あたりでも出生あたりでも多胎で有意に高頻度であった。その他、二分脊椎、大動脈の先天異常などいくつかの中分類項目に関しても多胎で有意に高頻度であった。

#### (2) 双胎ペアでの先天異常の一致率<sup>29,30)</sup>

これまで先天異常に関するふたご研究は児の卵性が確定しておらず、遺伝要因の影響(遺伝率)を正しく推定できないことが多かった。その中で、ARTで発生する多胎は大部分が二卵性双胎であり、その遺伝的類似度は一般の同胞(兄弟姉妹)と同じである点に注目した。症例リストは個人単位なので、入手可能な変数を基に双胎ペアを同定した。先天異常全体では、発端者一致率は9.6%と推定された。この値は、多因子遺伝を仮定したEdwardsの推定式から求めた同胞再発危険率(9.0%)と類似しており、先天異常は主に多因子性疾患であることが支持された。

個別の疾患で見ると、再発危険比が有意に高いのは「眼、耳、顔面及び頸部の先天奇形」「唇裂を伴う口蓋裂」であり、これらの疾患に強い家族集積性が示唆された。

### 5.3 統計データ以外の活用例

前述のように既存統計データの分析で、多胎児家庭の虐待リスクが高いことが分かった。そこで、虐待の要因を探る目的で多胎児家庭における虐待死亡事例の特徴を分析した<sup>31)</sup>。虐待死亡事例の報告書が「新聞報道等から抽出し、地方公共団体が把握した死亡事例と合わせて地方公共団体に詳細を調査した」ことを利用した。14児の多胎死亡事例について、複数の新聞記事やインターネットなどから詳細な情報を収集した結果、同数の事例を確認できた(厳密な照合はデータの性質上不可能である)。

分析の結果、双胎両児が虐待を受けた事例では、家族の機能不全が認められた。双胎の1児だけが虐待を受けた事例では、1児の障がい、成長・

発達の違い、愛情の偏りなどが見られた。虐待死亡に至った児の性別、両親の年齢、婚姻状況、母親の精神状態など大部分の項目で単胎児と多胎児の間に差はなかった。ただし、0か月死亡が多胎児で0%(8児中0児)である一方、単胎児では48%(210児中100児)であり有意差が見られた。多胎では出産前後を通じて、母児の入院期間が長くなりやすく、虐待を予防・抑制した可能性がある。また、多胎児家庭では単胎児家庭と比べて児の数が有意に多かった。子どもの数が多いこと自体が多胎育児を困難にし、虐待リスクを高める要因の一つだと言える。

多胎児家庭が単胎児家庭よりも虐待死亡事例の発生頻度が高い理由は、多胎児家庭に限定されない一般的な要因と、多胎児家庭に固有な要因が関係するといえる。

自験例を通じて様々な公表データの活用を具体的に述べた。質問紙調査や面接調査などで調査対象者から新たなデータを集めなくても、公表されたデータを用いて様々な知見が比較的短時間に、しかも安価に得られる。また、Excelだけで十分に分析可能な場合も多く、保健医療系の学部教育でも実施できる研究方法であると考えられる。通常の質問紙調査や面接調査などでは研究困難な知見もある。得られた結果は、学術面だけでなく、実践面(多胎児家庭の育児支援)でも重要なエビデンスとなる。この種のデータは継続的に収集可能であるため、モニタリングにも有効である。

## 6. まとめ

公表された既存データを活用した研究の課題点の多くは、データの妥当性や研究結果の解釈など、理論的な問題に付随する。教育的見地から見た場合には、公表されたデータの分析には利点も多い。国・都道府県レベルの様々なデータが公表されている。研究方法論の選択肢として、公表された既存データの有効活用は、十分に検討する価値があると思われる。

### 謝辞

研究アシスタントの大間敏美さんに多大なご協力を頂きました。

### 利益相反

なし



## 引用文献

- 1) Cwikel, J.G.: Methods of descriptive epidemiology. Social Epidemiology: Strategies for Public Health Activism. Columbia University Press, New York, 180-220, 2006.
- 2) Morgenstern, H.: Uses of ecologic analysis in epidemiologic research. Am J Public Health, 72(12), 1336-1344, 1982.
- 3) Walter, S.D.: The ecologic method in the study of environmental health. I. Overview of the method. Environ Health Perspect, 94, 61-65, 1991.
- 4) Walter, S.D.: The ecologic method in the study of environmental health. II. Methodologic issues and feasibility. Environ Health Perspect, 94, 67-73, 1991.
- 5) Susser, M.: The logic in ecological: I. The logic of analysis. Am J Public Health, 84 (5) , 825-829, 1994.
- 6) Susser, M.: The logic in ecological: II. The logic of design. Am J Public Health, 84 (5) , 830-835, 1994.
- 7) Morgenstern, H.: Ecologic studies in epidemiology: concepts, principles, and methods. Annu Rev Public Health, 16, 61-81, 1995.
- 8) Donne, J.: Ecological studies. Koepsell, T.D. and Weiss, N.S. eds.: Epidemiologic Methods: Studying the Occurrence of Illness. Oxford University Press, New York, 281-307, 2003.
- 9) Morgenstern, H.: Ecologic studies. Rothman, K.J., Greenland, S., Lash, T. eds.: Modern Epidemiology, Third edition. Lippincott Williams & Wilkins, Philadelphia, 511-531, 2008.
- 10) 尾島俊之: 既存統計資料の利用. 柳川洋編: 地域保健活動のための疫学 第2版, 日本公衆衛生協会, 53-63, 2006.
- 11) Lawson, A.B.: Statistical Methods in Spatial Epidemiology, Second edition. Wiley, New York, 2006.
- 12) Glenn, N.D.: Cohort Analysis, Second edition. SAGE Publications, London, 2005.
- 13) 佐藤博樹: JGSS の授業への利用: 日本社会学会での報告・議論を参考に. JGSS 研究論文集, 2, 233-236, 2003.
- 14) 浜松誠二: 公式統計調査とその二次分析. [http://www.tuins.ac.jp/~ham/tymhnt/analysis/b/b\\_seco.html#i3](http://www.tuins.ac.jp/~ham/tymhnt/analysis/b/b_seco.html#i3), access 2013/11/20
- 15) Barker, D.J., Osmond, C.: Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. Lancet, 1 (8489) , 1077-1081, 1986.
- 16) Barker, D.J.: Fetal Origins of Cardiovascular and Lung Disease. CRC Press, Florida 1-2, 2001.
- 17) Renaud, S., de Lorgeril, M.: Wine, alcohol, platelets, and the French paradox for coronary heart disease. Lancet, 339 (8808) , 1523-1526, 1992.
- 18) 厚生労働省告示第四百三十号: 国民の健康の増進の総合的な推進を図るための基本的な方針. [http://www.mhlw.go.jp/bunya/kenkou/dl/kenkounippon21\\_01.pdf](http://www.mhlw.go.jp/bunya/kenkou/dl/kenkounippon21_01.pdf), access 2013/9/19
- 19) Rose, G.: Sick individuals and sick populations. Int J Epidemiol, 14 (1) , 32-38, 1985.
- 20) Marmot, M.: Economic and social determinants of disease. Bull World Health Organ, 79 (10) , 988-989, 2001.
- 21) World Health Organization: Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health. World Health Organization, Geneva, 2008.
- 22) Ooki, S.: Effect of maternal age and fertility treatment on the increase in multiple births in Japan: vital statistics, 1974-2009. J Epidemiol, 21(6), 507-511, 2011.
- 23) Ooki, S.: Estimation of the contribution of assisted and non-assisted reproductive technology fertility treatments to multiple births during the past 30 years in Japan: 1979-2008. Twin Res Hum Genet, 14 (5) , 476-483, 2011.
- 24) Ooki, S.: Theoretical model of the relationship between single embryo transfer rate and multiple pregnancy rate in Japan. J Pregnancy, doi: 10.1155/2012/620753, 2012.
- 25) Ooki, S.: The effect of an increase in the rate of multiple births on low-birth-weight and preterm deliveries during 1975-2008. J Epidemiol, 20 (6) , 480-488, 2010.
- 26) Ooki, S.: Fatal child maltreatment associated with multiple births in Japan: nationwide data between July 2003 and March 2011. Environ Health Prev Med, 18 (5) , 416-421, 2013.
- 27) Ooki, S.: Birth defects in singleton versus multiple ART births in Japan (2004-2008) . J Pregnancy, doi: 10.1155/2011/285706, 2011.
- 28) Ooki, S.: Birth defects after assisted reproductive technology in Japan: comparison between multiples and singletons, 2004-2009. Reprod Sys Sexual Disorders, S:5, doi:10.4172/2161-038X.S5-003, 2012.

- 29) Ooki, S.: Nationwide study of assisted reproductive technology and multiple births with accompanied birth defects. Sanger, I. eds.: Advances in Reproductive Technology Research. Nova Science Publishers, New York, 1-70, 2013.
- 30) Ooki, S.: Concordance rates of birth defects after assisted reproductive technology among 17,258 Japanese twin pregnancies: a nationwide survey, 2004-2009. J Epidemiol, 23 (1) , 63-69, 2013.
- 31) Ooki, S.: Characteristics of fatal child maltreatment associated with multiple births in Japan. Twin Res Hum Genet, 16 (3) , 743-750, 2013.

## On the Effective Use of Open Data

Syuichi OOKI, Kiyomi HIKO

### Abstract

We discussed the effective use of open data (that is, data not restricted to statistical data), which are relatively available. Examples of the effective use of open data are descriptive epidemiologic studies and ecologic/ecological studies in epidemiology. Such studies use population/aggregative statistical data. Prior discussions of the drawbacks of the studies using aggregative data noted methodological or theoretical problems. But there are many advantages of using open data, including not only statistical data, but many other types of data, in educational settings, especially for undergraduate education in healthcare science. Studies using open data can be completed relatively quickly and at low cost. Research examples of hypothesis generation using open data, which became the preface of future big studies in health science were shown. Moreover, future effective use of open data, for example to consider the correction of health inequalities was also shown. In addition, there are many open data to be used for a wide variety of research purposes. Here we also introduced research examples of the use of open data. The basic concepts are applicable to other research themes in health-related sciences.

**Key words** open data, effective use, ecologic/ecological study, hypothesis generation, undergraduate education in healthcare science