厚生労働科学研究費補助金
（政策科学総合研究事業（統計情報総合研究））
分担研究報告書

市区町村別性年齢階級別人口の線形補間について

研究協力者　　福井敬祐　大阪府立成人病センターがん予防情報センター　研究員

研究要旨

市区町村別地理的剥奪指標を用いた全死亡・主要死因別の年齢調整死亡率を算出するためには、市区町村別の性別・年齢 5 歳階級別の人口が必要となる。しかしながらこれらの人口データは国勢調査実施年のものしか提供されていない。そこで、本研究における人口動態統計の分析対象期間である 1985～2014 年における国勢調査実施年以外の年における市区町村別性年齢階級別人口の線形補間の実施手順および課題についてまとめた。対象期間の間、市区町村は合併・分割・分割後合併が行われている。Web にて桐村らが提供する「Municipality Map Maker ウェブ版 市区町村区域の GIS データ生成ツール」を用いて、市区町村構成の変遷に対応した。市区町村構成の変遷パターンに応じ、2 時点の市区町村における性別・年齢 5 歳階級別人口の線形補間を行った。本手法による補間において以下の課題が残った。①東京都三宅村の人口が 0 になる、②外挿した場合の人口が負の値になるところがある、③直近の国勢調査実施年の間で線形補間したが、対象期間共通の市区町村区分で統一し、線形補間した方がよいか、④基準日の詳細設定、⑤面積による重み付けの問題などである。今後、上記課題を解決し、より精緻な人口データセットを作成する必要がある。

A．研究目的

国勢調査における市区町村別の人口データは 1985 年, 1990 年, 1995 年, 2000 年, 2005 年, 2010 年までの 5 年毎のものとなっている. より安定した分析を行うためには各 5 年の間のデータを補間し利用することが考えられる. 補間法として一般的な線形補間は隣接する 2 時点の対応するデータを用いて行われるが, 市区町村別に着目すれ場合には合併や分割など影響を考慮する必要がある. 本報告は国勢調査から得られた人口データにおいて, 合併を考慮した上で線形補完を行う方法についてまとめたものである.

B．研究方法

1．具体例

図 1 は熊本県球磨郡上村, 免田町, 岡原村, 須恵村, 深田村の 5 町村の合併の 2000 年から 2005 年の変遷を表している. 5 町村が 2003 年 10 月 1 日を以てあさぎり町として合併していることがわかる. 国勢調査は 5 年毎に行われるため, 今, 2000 年と 2005 年の人口データのみが得られている. 図内

の5町村は2005年時点では合併により存在しないため，2000年と2005年の間でそのまま線形補間を行うことができない．このような場合においては2000年を合併後の市区町村に，もしくは2005年を合併前の市区町村に作り変え，仮想的に対応した2時点を作り出すことで線形補間を行う．図2は2000年時点の5町村を2000年時点で仮想的にあさぎり町に合併することで，2005年時点と対応可能にし，線形補間を行うイメージ図を表している．仮に，2000年時点の上村，免田町，岡原村，須恵村，深田村の人口がそれぞれ，45千人，100千人，30千人，20千人，15千人であったとすれば，2000年時点で仮想的に作成されたあさぎり町の人口は5町村の人口を足し合わせた210千人である．さらに2005年時点のあさぎり町の人口が135千人であったと仮定すると，仮想的に作成された2000年時点の人口210千人との線形補間により2001年，2002年，2003年，2004年の人口はそれぞれ195千人，180千人，165千人，150千人と計算される．

2．使用したデータ

人口データ

国勢調査より入手した1985年，1990年，1995年，2000年，2005年，2010年の市区町村・性・年齢階級別人口データ

市区町村変遷対応表

桐村らが提供する「Municipality Map Maker ウェブ版 市区町村区域のGISデータ生成ツール」[1]より以下の5つのcsvファイルを作成・入手した．

① 1985年10月1日時点から1990年10月1日時点への市区町村対応表
② 1990年10月1日時点から1995年10月1日時点への市区町村対応表
③ 1995年10月1日時点から2000年10月1日時点への市区町村対応表
④ 2000年10月1日時点から2005年10月1日時点への市区町村対応表
⑤ 2005年10月1日時点から2010年10月1日時点への市区町村対応表

図3は入手した④の市区町村対応表の一例である．市区町対応表を用いて，2000年時の住所区分けを表す住所コード(JISCODE1)に2005年時の住所区分けを表す住所コード(JISCODE2)を対応させる(紐付ける)ことができる．

3．補間方法

線形補間は市区町村変遷対応表のJISCODE2をJISCODE1に紐付ける(最新年の市区町村分けに対応させる)ことで線形補間に用いる1対1の対応を作成した後に行う．紐付けの方法は市区町村の変遷パターン（エラー! 参照元が見つかりません。）に大きく依存するため，線形補間の方法についてもこのパターンに沿って説明する．なお2時点 t_1, t_2 年（$t_1 < t_2$）に対応する人口を y_1, y_2 としたとき，区間 $[t_1, t_2]$ 内の任意の時点 t の人口 y は

$$y = y_1 + \frac{y_2 - y_1}{t_2 - t_1} \times (t - t_1),$$

によって計算される．

(1) 変化しない

時間が経過しても市区町村が変化しない場合にはJISCODE2をJISCODE1に紐付け，2時点の人口を使用して線形補間を行う．

(2) 合併

合併が起きた場合には，対応表を元にJISCODE2(合併後住所)をJISCODE1(合併前住所)に紐付けしたあと，JISCODE2が同じ市区町村の人口を合算することで仮想的に最新年と同じ市区町村分けを作成し線形補間を行う(図 5).

(3) 分割

分割の場合には，JISCODE2をJISCODE1に紐付けたあと，市区町村変遷対応表内にあるWEIIGHを用いて最新年の市区町村分けに重み付けで人口を分割する．その後，対応した2時点間でそれぞれ線形補間を行う(図 6).

(4) 分割後合併

分割後合併の場合にはJISCODE2をJISCODE1に紐付けたあと，市区町村変遷対応表内にあるWEIIGHを用いて最新年の市区町村分けに重み付けで人口を分割する．その後，紐付けられたJISCODE2が同じ市区町村の人口を合算し仮想的な人口を作成し線形補間を行う(図 7).

(倫理面への配慮)

本研究に用いた資料は全て公開データに基づいているため、倫理面において問題になることはない。

C・D．研究結果および考察

本手法による補間では以下の課題が残っている。

(ア) 三宅村の人口

平成12年国勢調査の際，三宅島噴火により，全島民が島外へ避難したことにより，東京都三宅村の人口は0となっている．この場合どのように取り扱うのか.

(イ) 外挿した場合の取り扱い

外挿において地区・年齢階級によっては人口が負の値になる地区も存在する．単純に0と置き換えてよいか.

(ウ) 線形補間の区間（基準点をどうするか）

線形補間は国勢調査が行われた5年間で直近分の市区町村区分けに紐づけして行っているが実際には全期間で統一した方がよいのではないか。

(エ) 使用する市区町村変遷対応表基準日

市区町村変遷対応表は現在取得を各国政調査が行われた年の10月1日を基準日としているが，より細かく基準日を設定する必要がある可能性がある.

(オ) WEIGHTの使用について

分割の際に利用する市区町村変遷対応表の重み(WEIGHT)は土地面積比により作成されているが，面積が大きいが人口は少ないという土地に分割を行った場合には，実際の人口から大きく乖離し，線形補間が不安定になる場合がある.

静岡県浜松市の例

静岡県浜松市は2009年9月1日に中区，東区，西区，南区，北区，浜北区，天竜区の7区に分割された．2010年における各区の人口割合は．浜松市人口割合

図 9. . 浜松市人口割合

図 9の通りであるが，図 9にある通り本来人口が最も少ない天竜区に対するWEIGHTが最大となって

いる. 2005 年時点の浜松市の人口を 2010 年の区分けに対応させるための WEIGHT を用いた分割時に天竜区の人口は 2005 年時点の人口 399 千人に 0.626657 をかけたおよそ 250 千人と算出される. しかし, 2010 年時点の人口は 16 千人とその差が大きい, そのため補間も実際の人口推移とは乖離するものとなってしまう.

E. 結論

本報告書に記載した方法で作成した補間人口データは 1985 年, 1990 年, 1995 年, 2000 年, 2005 年, 2010 年の国勢調査のデータの線形補間法について記述した. 作成したデータは 1986 年～1989 年, 1991 年～1994 年, 1996 年～1999 年, 2001 年～2004 年, 2006 年～2009 年の単年データである. これらのデータはそれぞれ直近の国勢調査に対応する市区町村区分けに変換しているのみ(例えば, 1986 年～1989 年であれば 1990 年に変換)であり, 全期間を通して同じ市町村区分けを利用している訳でないことに注意されたい. 今後は 2000 年などのようにある特定の時点の市区町村分けに変換し, 全期間で比較・対応可能なデータの作成を行いたい.

F. 健康危険情報

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

引用文献

[1] 桐村 喬, 「Municipality Map Maker ウェブ版 市区町村区域の GIS データ生成ツール」, 〈http://www.tkirimura.com/mmm/〉 (参照 2016 年 4 月 2 日).

2000年	2001年	2002年	2003年		2004年	2005年
			~10/1	10/1~		
上村	上村	上村	上村	あさぎり町	あさぎり町	あさぎり町
免田町	免田町	免田町	免田町			
岡原村	岡原村	岡原村	岡原村			
須恵村	須恵村	須恵村	須恵村			
深田村	深田村	深田村	深田村			

図 1.　2000 年から 2005 年間で行われた熊本県球磨郡あさぎり町合併の変遷

図 2. 2000 年から 2005 年における合併を考慮した線形補間のイメージ

NO	DATE1	JISCODE1	PNAME1	GNAME1	CNAME1	WEIGHT	DATE2	JISCODE2	PNAME2	GNAME2	CNAME2
469	20001001	43502	熊本県	球磨郡	上村	1	20051001	43514	熊本県	球磨郡	あさぎり町
470	20001001	43503	熊本県	球磨郡	免田町	1	20051001	43514	熊本県	球磨郡	あさぎり町
471	20001001	43504	熊本県	球磨郡	岡原村	1	20051001	43514	熊本県	球磨郡	あさぎり町
472	20001001	43508	熊本県	球磨郡	須恵村	1	20051001	43514	熊本県	球磨郡	あさぎり町
473	20001001	43509	熊本県	球磨郡	深田村	1	20051001	43514	熊本県	球磨郡	あさぎり町

図 3.　市区町村変遷対応表の例

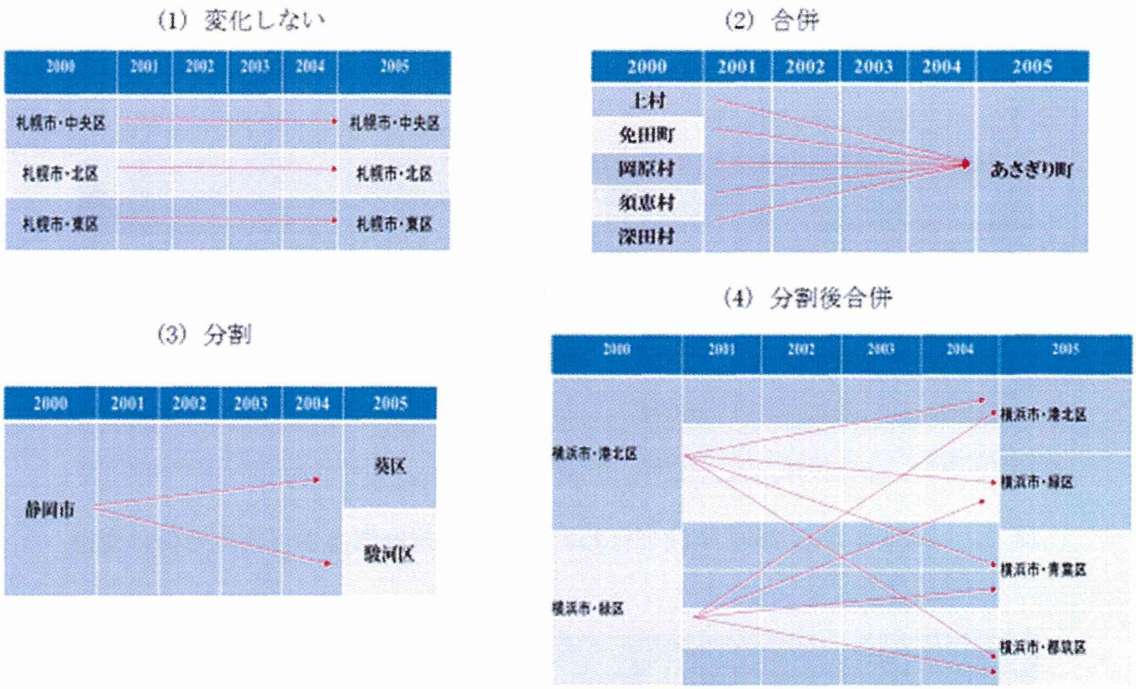


図 4. 市区町村変遷パターン

① 市区町村変遷対応表を元にあさぎり町を2000年時点の上村・免田町・岡原村・須恵村・深田村に紐付け，人口を合算することで2000年時点の仮想的なあさぎり町人口を作成。

② ①にて作成した2000年のあさぎり町人口と2005年のあさぎり町人口で線形補間．つまり，

$$x\text{年の人口} = 210 + \frac{135 - 210}{2005 - 2000} \times (x - 2000), \qquad x = 2001, \ldots, 2004$$

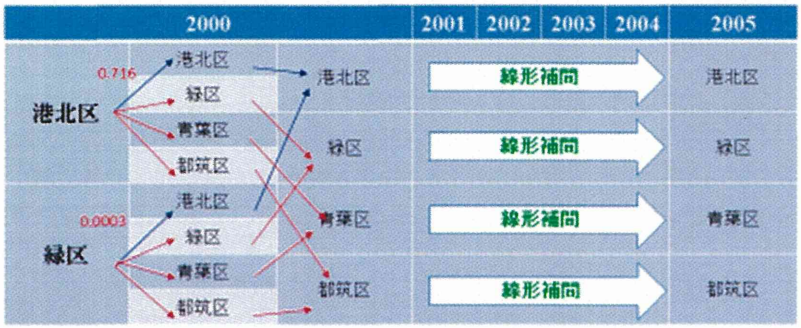
図 5. 市区町村変遷・合併の例．かっこ内は仮想的な人口（千人）を表す

NO	DATE1	JISCODE1	PNAME1	GNAME1	CNAME1	WEIGHT	DATE2	JISCODE2	PNAME2	GNAME2	CNAME2
448	20001001	22201	静岡県		静岡市	0.936	20051001	22101	静岡県	静岡市	葵区
448	20001001	22201	静岡県		静岡市	0.064	20051001	22102	静岡県	静岡市	駿河区

① 市区町村変遷対応表を元に葵区・駿河区を2000年時点の静岡市と紐付け，対応表内のWEIGHTを用いて人口を分割
② 分割させた2000年時点の葵区・駿河区人口と2005年時点の葵区・駿河区人口でそれぞれ線形補間

図 6. 分割時に関する市区町村変遷対応表と市区町村変遷・分割の例

NO	DATE1	JISCODE1	PNAME1	GNAME1	CNAME1	WEIGHT	DATE2	JISCODE2	PNAME2	GNAME2	CNAME2
360	19901001	14109	神奈川県	横浜市	港北区	0.71641	19951001	14109	神奈川県	横浜市	港北区
360	19901001	14109	神奈川県	横浜市	港北区	0.000658	19951001	14113	神奈川県	横浜市	緑区
360	19901001	14109	神奈川県	横浜市	港北区	0.000889	19951001	14117	神奈川県	横浜市	青葉区
360	19901001	14109	神奈川県	横浜市	港北区	0.282043	19951001	14118	神奈川県	横浜市	都筑区
361	19901001	14113	神奈川県	横浜市	緑区	0.000319	19951001	14109	神奈川県	横浜市	港北区
361	19901001	14113	神奈川県	横浜市	緑区	0.332717	19951001	14113	神奈川県	横浜市	緑区
361	19901001	14113	神奈川県	横浜市	緑区	0.46172	19951001	14117	神奈川県	横浜市	青葉区
361	19901001	14113	神奈川県	横浜市	緑区	0.205244	19951001	14118	神奈川県	横浜市	都筑区



①市区町村変遷対応表を元に2000年時点の区を，対応表内のWEIGHTを用いて人口を分割．
②分割させた2000年時点の区を2005年時点の区分けを用いて合算
③合算させた人口を2005年人口と対応させ線形補間

図 7. 分割後合併時に関する市区町村変遷対応表と市区町村変遷・分割の例

市	区名	2010年総人口
浜松市	中区	119044
	東区	63053
	西区	56162
	南区	51420
	北区	46260
	浜北区	44915
	天竜区	16292

図 8. 浜松市人口割合

DATE1	JISCODE1	GNAME1	CNAME1	WEIGHT	DATE2	JISCODE2	GNAME2	CNAME2
20051001	22202		浜松市	0.029425	20101001	22131	浜松市	中区
20051001	22202		浜松市	0.03097	20101001	22132	浜松市	東区
20051001	22202		浜松市	0.056619	20101001	22133	浜松市	西区
20051001	22202		浜松市	0.031523	20101001	22134	浜松市	南区
20051001	22202		浜松市	0.180607	20101001	22135	浜松市	北区
20051001	22202		浜松市	0.044198	20101001	22136	浜松市	浜北区
20051001	22202		浜松市	0.626657	20101001	22137	浜松市	天竜区

図 9. 市町変遷対応表（一部抜粋）

厚生労働科学研究費補助金
（政策科学総合研究事業（統計情報総合研究））
分担研究報告書

Probabilistic linkage を用いた大規模公的統計データベースの活用に関する研究

研究協力者　福井敬祐　大阪府立成人病センターがん予防情報センター　研究員
研究分担者　近藤尚己　東京大学大学院医学系研究科　准教授

研究要旨

我が国の人口動態特殊報告を用いた職業・産業別死亡情報は、死亡時職業による分類であるため、実際に長く従事した職業や産業による影響を正確に計ることができないことが懸念されてきた。国勢調査における職業・産業情報と人口動態統計を突合することができれば、その問題点の解決につながる。個人識別番号によるリンケージが不可能な大規模統計データ同士を、個人単位でリンケージする手法として、複数の変数同士の一致確率を統計的に算出して行う probabilistic record linkage の理論と実際について整理した。Probabilistic record linkage においては、リンケージを行うデータベース間での共通のマッチング変数が重要となる。しかしながら、本研究で想定している国勢調査および人口動態統計の二次利用データで入手可能な変数では、あまり正確にリンケージできない可能性がある。また、住所等の情報の入り方の問題もあるため、かなりの作業量を要することが予想された。将来的には Probabilistic record linkage に頼る必要がないよう個人識別番号による突合が可能となるような基盤整備が必要であることが示唆された。

A．研究目的

我が国の人口動態特殊報告を用いた職業・産業別死亡情報は、死亡時職業による分類であるため、実際に長く従事した職業や産業による影響を正確に計ることができないことが懸念されてきた。国勢調査における職業・産業情報と人口動態統計を突合することができれば、その問題点の解決につながるが、現状では氏名・生年月日・住所等、個人をつなぐ情報を使うことができないため、完全なリンケージデータの作成は困難である。

二次利用申請により提供可能である限られた個人属性を用いて、国勢調査データと人口動態調査データをリンケージすることが可能であるかどうか、諸外国で活用されているProbabilistic Record Linkage法について紹介し、我が国における活用可能性を検討する。

B．研究方法

Probabilistic Record Linkageを含む一般的なRecord Linkage手法について、その実行方法を調査し、整理した。特に、Probabilistic Record Linkageについては理論的な背景を中心にまとめ，実行可能なソフト

ウェア等についても紹介した。

（倫理面の配慮）

本報告では実際のデータを用いていないため、倫理面での問題は生じない。

C．研究結果

1. Record Linkage の概要

今、2 つのデータセット A、B を考える。データセットを構成する各データはレコードと呼ばれ、各データセットの中からペアとなるレコードを作成することが Record Linkage の目的である。Linkage に使用する各データセットに共通した変数を Matching Variables (MV)と呼び、MV に基づいて Linkage を行う。例えば、Record Linkage 法として最も単純な Deterministic Record Linkage においては MV の値が同じレコードを Linkage する。図 10 はデータセット A とデータセット B を Record Linkage を行う場合の例である。ここでは ID と名前を MV として、レコード *a* とレコード *b* を Linkage している。

一般的な Linkage の手順は研究毎に様々であるが（例えば、[1, 2, 3]など）それらの方法は簡単に以下の手順に包含することができる (図 11)

① **Data cleaning and standardization**

データの整形にあたる部分である。Linkage を行うデータの誤った入力の修正や、データ間で異なった入力値を共通なものへと変換し、標準化する。

② **Blocking**

各データの間での Linkage 作業における比較数を削減するため、データを Block と呼ばれるいくつかのグループに分割する。Linkage は対応する Block 間のみで行われる。例えば、性別という変数を共通に持つデータ A とデータ B をリンケージする場合には、あらかじめ男性の Block、女性の Block というように Blocking を行う。このようにすることで、Linkage の際には男性の Block 同士、女性の Block 同士のみを比較すればよく、計算量を削減することができる。

③ **Linkage**

Blocking によって作成された Block 間で MV を基にして Linkage 作業を行う。Linkage の方法としては Deterministic な方法と Probabilistic な方法がある。

Deterministic 法は MV が一致するレコード同士をペアとして Linkage する方法であり、Probabilistic 法は MV を基にして作った Weight を使って Linkage を行う方法である。Deterministic 法は非常に単純な方法であり、一般的に MV を種別化・順序化し、データセット同士で結合作業をすればよい。しかし、例えば MV がユニークでない値を含む場合や欠損値が存在する場合などには Linkage の精度は大きく低下する。一般的に Record Linkage を行うデータ量は巨大になりがちであり、データの質の担保が困難な場合が多い。そのため、Deterministic 法を用いた Record Linkage では期待した通りの結果が得られないことが多い。

Probabilistic 法は MV を基に作成された重み（Weight）を利用して Linkage を行う。Weight は Link させたレコードが真のペアであるか否かの確率を反映しており、MV の値を Exact に用いないた

め、データの質が低い場合でも使用できる可能性がある.

④ **Clerical review**

Linkage 作業で Linked もしくは Non-Linked に判別されないようなデータを人為的に判別することを指す。 また、Linkage 作業に用いた閾値などのパラメータ設定やソフトウェアの実行が正しいものであったのかを判別することも含む。

⑤ **Evaluating data quality**

省略

2. Probabilistic Record Linkage について

Probabilistic Record Linkage の基本的な考え方は[4]により提案された。今、$\boldsymbol{A}$、$\boldsymbol{B}$ を Linkage 対象のデータ、$a \in \boldsymbol{A}$、$b \in \boldsymbol{B}$を任意のそれぞれのレコードとする。このとき、直積集合$\boldsymbol{A} \times \boldsymbol{B} = \{(a,b)|a \in \boldsymbol{A},\ b \in \boldsymbol{B}\}$の２つの部分集合を

$$\boldsymbol{M} = \{(a,b)|a = b,\ a \in \boldsymbol{A},\ b \in \boldsymbol{B}\},$$
$$\boldsymbol{U} = \{(a,b)|a \neq b,\ a \in \boldsymbol{A},\ b \in \boldsymbol{B}\},$$

とすれば、$\boldsymbol{M}$は真に Link 関係にあるレコードの集合、$\boldsymbol{U}$は真には Link 関係にないレコードの集合を表す。また、$\boldsymbol{t}(a,b)$をレコードa,bの一致度を測る一致度ベクトルとし、レコードa,bが真に Link 関係にあるとき、一致度ベクトルが$\boldsymbol{t}$となる確率を$m(\boldsymbol{t})$と定義する。すわわち、

$$m(\boldsymbol{t}) = P(\boldsymbol{t}(a,b) = \boldsymbol{t}|(a,b) \in \boldsymbol{M}).$$

同様に、レコードa,bが真に Link 関係にないときに、一致度ベクトルが$\boldsymbol{t}$となる確率を$u(\boldsymbol{t})$で定義する。すなわち、

$$u(\boldsymbol{t}) = P(\boldsymbol{t}(a,b) = \boldsymbol{t}|(a,b) \in \boldsymbol{U}).$$

このとき、[4]は次の Weight を用いて Link か否かを決定する方法を提案している。

$$w(\boldsymbol{t}) = \log\frac{m(\boldsymbol{t})}{u(\boldsymbol{t})}.$$

実際には、一致度ベクトルが取り得る値$\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k$の全てに対して$w(\boldsymbol{t}_1), \ldots, w(\boldsymbol{t}_k)$を計算し、あらかじめ設定した閾値と比較するという方法をとる。上記の Weight の計算には、レコード内の角変数に独立性を仮定し、EM アルゴリズムを利用する方法が提案されている(参照[5])。

3. 応用ソフトウェア等

Probabilistic Record Linkage を行うことができるソフトウェアは有償・無償のもの含めて様々開発がされている。また一般的な統計ソフトウェアのパッケージとして提供されているなど、導入しやすい。例えば、National Program of Cancer Registries (NPCR) によって開発・提供されている無償の Record Linkage ソフトウェア Link Plus (図 3)は GUI ユーザーインターフェイスで直感的な操作が可能であり、比較的簡単に Record Linkage が可能である。他にも Record Linkage を行える GUI ベースのソフトとしては、無償のものでは D-Dupe、DuDe、Merge Tool Box などがある。その他のソフトについては[6]を参照されたい。また、統計解析ソフトウェアの R 言語における RecordLinkage パッケージは多少のプログラミング知識を有するが、多量なデータを自動で Linkage したい場合や Linkage したデータを直接分析する必要がある場合などに有用である。

これらのソフトウェアを用いる上での注意点としては，無償版のソフトウェアのほとんどが日本語に完全に対応しいるわけではないということである。そのため、入力値や MV として日本語が含まれるデータの Linkage の際にはその精度に対して十分な注意が必要である。

4. 実適用について

Probabilistic Record Linkage の基礎理論は上述したとおり、[4]により提案されており、その歴史は長く、海外を中心に活用されている（例えば、[7,8,9]など）。一方で、日本における適用例はまだ少なく、[10]においては、Record Linkage を必要とする分野がそれぞれ領域固有の知識を必要とするために、学術的な一般化が難しかったこと、 我が国では戸籍制度が整備されており、国勢調査等で人物同定の必要性がほとんどなかったという社会背景をその理由として挙げている.

D．考察

個人識別番号によるリンケージが不可能な大規模統計データベースを、個人単位でリンケージする手法として、複数の変数同士の一致確率を統計的に算出して行うProbabilistic Record Linkageの理論と実際について調べた。

Record Linkage法は上記に上げた手順①～⑤の作業を必要とするが、応用ソフトウェアではこれらの作業を支援することができるものがほとんどである。 しかし，Linkageに有効なMVとして用いられがちな氏名や住所等の情報は基本的に日本語での入力が行われるため，ソフトウェアでMVとして正確に作用しない可能性がある.

本研究でLinkageを考える人口動態統計と国勢調査データを使用する場合、現行でLinkageに使用可能な変数は以下の通りである。

人口動態統計

- 事件簿番号
- 性別
- 生年月日
- 死亡年月日
- 死亡した人の住所
  死亡票：市区町村コード
  オンライン報告分：詳細住所

国勢調査

- 性別
- 生年月
- 調査区番号
- 市区町村（町丁目）情報

氏名や個人識別番号などの利用が困難である我が国の現状においては、利用可能な変数は有用なMVとなりにくく、ユニークでない組み合わせがかなり存在するリンケージデータとなる可能性が高い。住所情報に関しても共通のコード化などの工夫が必要となり、かなりの作業量を要することが想定される。

相当な作業量を要する上に、そのリンケージデータの精度が低いことが想定されるため、将来的には、北欧諸国や英国、米国のように、個人識別番号の整備を経て、各種公的統計のリンケージを公的機関が行い、個人識別可能な情報を削除した匿名化データを利用者に提供する仕組みが必要であると考える。

E．結論

Probabilistic Record Linkageの実行においては、現状の日本の国勢調査と人口動態統計データでは有用な共通のマッチング変数が利用可能でないため、精度の低いリンケージデータとなる可能性がある。Probabilistic Record Linkageにおける理論やソフトウェアの整備が進む一方で、得られる結果の整合性を考慮すれば、将来的には各種統計データベース間での共通個人識別番号の整備およびその活用について、検討していく必要がある。

F．健康危険情報

G．研究発表
1．論文発表
2．学会発表

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

引用文献

[1] Gu, L., Baxter, R., Vickers, D. and Rainsford, C. (2003). Record linkage: Current practice and future directions. *CSIRO Mathematical and Information Sciences Technical Report,* **3**, 83.

[2] Elfeky, M. G., Verykios, V. S. and Elmagarmid, A. K. (2002). TAILOR: A record linkage toolbox. In Data Engineering, 2002. Proceedings. 18th International Conference on (pp. 17-28). IEEE.

[3] Lee, M. L., Ling, T. W. and Low, W. L. (2000). IntelliClean: a knowledge-based intelligent data cleaner. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 290-294). ACM.

[4] Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328), 1183-1210.

[5] Bauman, G. J. (2006). Computation of weights for probabilistic record linkage using the EM algorithm.

[6] Christen, P., (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Springer Science and Business Media.

[7] Whop, L. J., Diaz, A., Baade, P., Garvey, G., Cunningham, J., Brotherton, J. M., ... & Moore, S. P. (2016). Using probabilistic record linkage methods to identify Australian Indigenous women on the Queensland Pap Smear Register: the National Indigenous Cervical Screening Project. BMJ open, 6(2).

[8] Kesinger, M. R., Kumar, R. G., Ritter, A. C., Sperry, J. L., & Wagner, A. K. (2016). Probabilistic Matching Approach to Link Deidentified Data from a Trauma Registry and a Traumatic Brain Injury Model System Center. American Journal of Physical Medicine & Rehabilitation.

[9] Adam, M., Kuehni, C. E., Spoerri, A., Schmidlin, K., Gumy-Pause, F., Brazzola, P., ... & Zwahlen, M. (2015). Socioeconomic Status and Childhood Leukemia Incidence in

Switzerland. Frontiers in oncology, 5.

[10] 相澤彰子, 高須淳宏, 大山敬三, & 安達淳. (2004). 異種データベース間でのレコード照合に関する研究動向. NII journal, (8), 43-51.

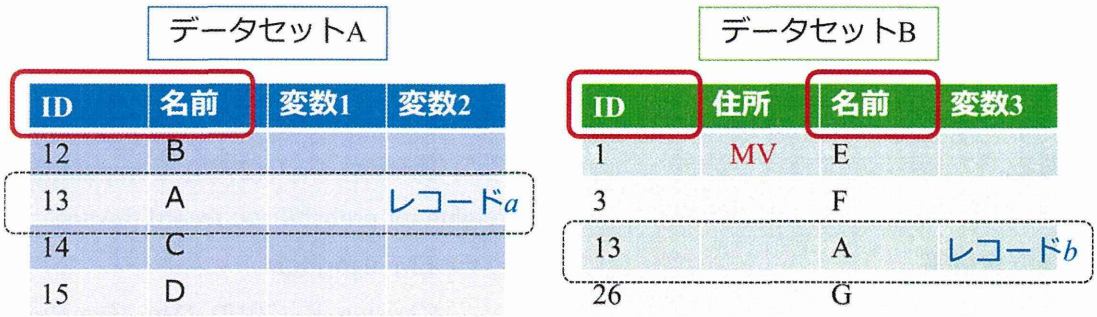


図 10 **Record Linkage** の例

データセット **A** 内のレコード **a** とデータセット **B** 内のレコード **b** を **ID** と名前を基に **Linkage**

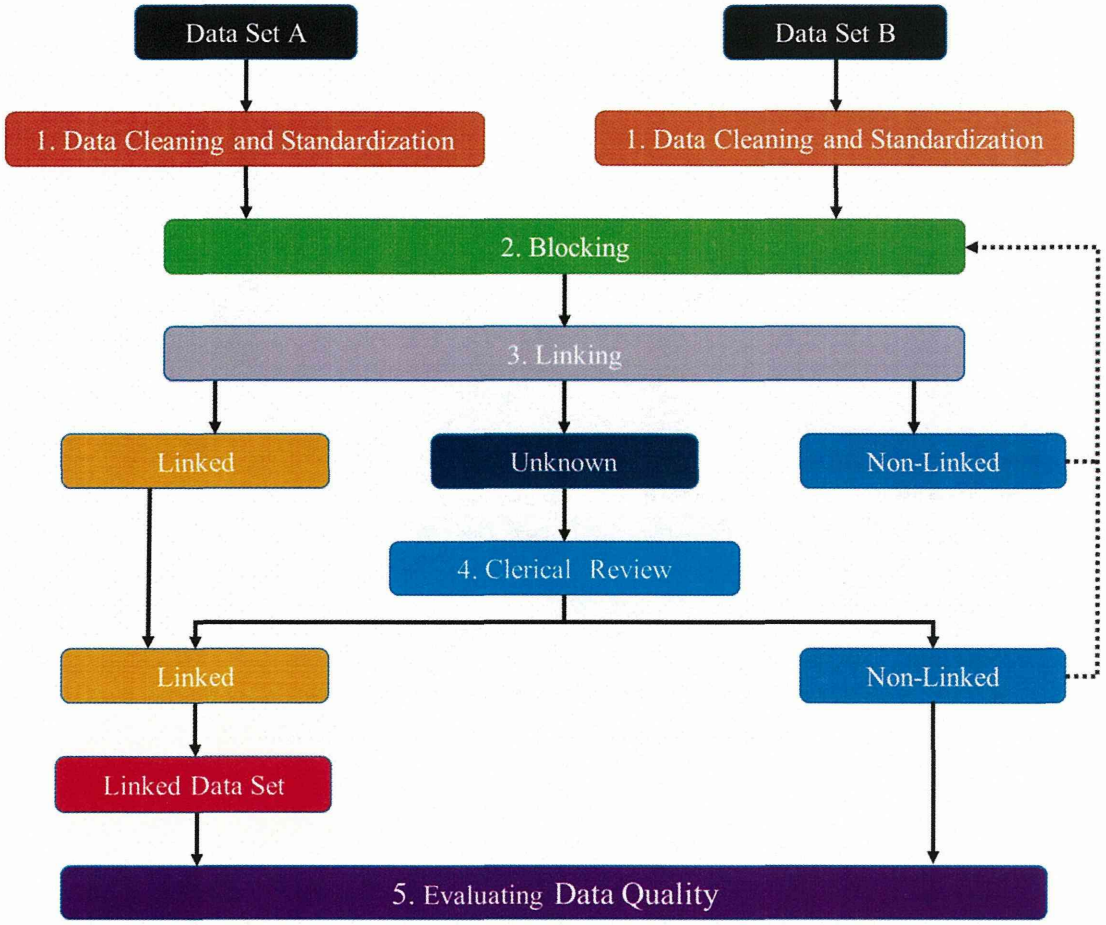


図 11 Record Linkageのフロー図

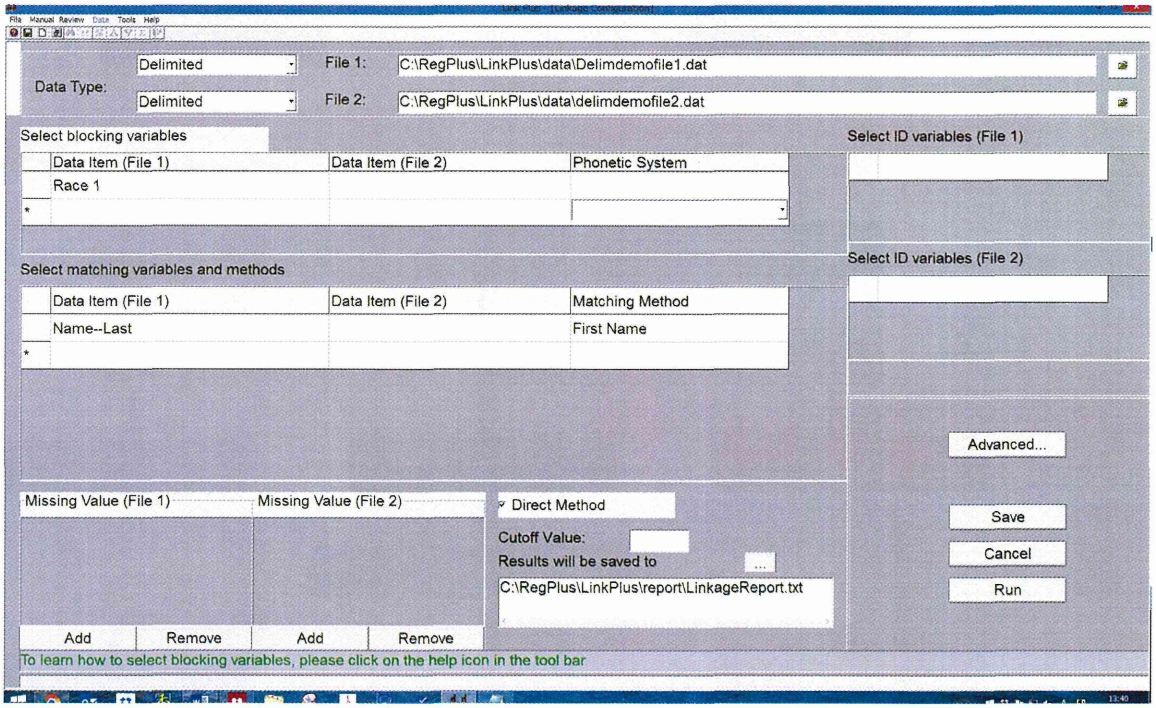


図 12 Link Plus のインターフェイス画面

厚生労働科学研究費補助金
（政策科学総合研究事業（統計情報総合研究））
分担研究報告書

がん進行度別罹患率における社会経済格差

研究代表者　伊藤ゆり　　大阪府立成人病センター　がん予防情報センター　主任研究員
研究分担者　近藤尚己　　東京大学大学院医学系研究科　准教授
研究分担者　中谷友樹　　立命館大学文学部（立命館大学歴史都市防災研究所 兼任）教授
研究協力者　米島万有子　立命館大学衣笠総合研究機構　専門研究員
研究協力者　中山富雄　　大阪府立成人病センター　がん予防情報センター　疫学予防課長
研究協力者　福井敬祐　　大阪府立成人病センター　がん予防情報センター　研究員

研究要旨

既存統計資料のひとつであるがん登録資料を用いて、がん罹患率における社会経済格差のトレンドを評価することを目的とする。大阪府がん登録資料より、大阪府において 1993-2004 年に診断された胃、大腸、肺、乳房、子宮頸部、前立腺のがん患者の情報を入手し、居住地（町字単位）を基に地理的剥奪指標（Areal Deprivation Index : ADI）を付与した。ADI は数値が大きいほど地域の剥奪度が高い、つまり社会的に不利な経済状況にある人々の割合が高いという指標で、各地域の人口で重み付けし 5 分位に分けた。がん罹患率は検診受診等の予防行動に影響を受けるため（検診受診率の高い地域の罹患率が高い）、診断時の進行度別（早期がん：上皮内＋限局／進行がん：領域＋遠隔転移）の年齢調整罹患率を ADI ごとに算出した。部位・性別・診断時期別に、分散重み付き最小二乗法により ADI と進行度別罹患率の関連を分析した。前立腺がん以外の全ての部位で、進行がんの罹患率は ADI の高い地域ほど高かった。ADI 第 5 分位と第 1 分位間の進行がんの罹患率差は、男女とも 1993-1998 年診断の肺がんが最大で、その差は 1999-2004 年には統計的有意に減少した。他の部位の進行がん罹患率においても、統計的有意には拡大しなかった。一方、早期がんの罹患率は、男性で前立腺、胃、大腸において、ADI の高い地域ほど罹患率が低かった。その傾向は前立腺で顕著であった。女性では子宮頸がんにおいて、ADI の高い地域の罹患率が高かった。格差の縮小が最も大きかったのは、男性の胃・大腸の進行がんであったが、この傾向は女性ではみられなかった。本研究により明らかとなった進行度別がん罹患率の社会経済格差には、喫煙やハイリスクな性行動など、がん発症のリスクとなりうる行動の違いやがん検診の受診率の違いなどが影響していると考えられる。

A．研究目的

平成 28 年 1 月 1 日より、がん登録等推進法に基づき、全国がん登録制度が開始した。地域がん登録資料は 1960 年代から一部の府県において収集されている既存統計資料であり、今後ますます健康政策への活用が求められている。健康格差の縮小が健康日本 21 の目標に掲げられ、健康格差の評価が求められているが、地域がん登録資料を用いた健康格差のモニタリング体制は十分ではない。本研究は既存統計資料である大阪府がん登録資料を用いて、がん罹患率における社会経済格差のトレンドを評価することを目的とする。

B．研究方法

大阪府において 1993-2004 年に診断された胃、大腸、肺、乳房、子宮頸部、前立腺のがん患者の情報に対し、居住地（町字単位）を基に地理的剥奪指標（Areal Deprivation Index : ADI）を付与した。ADI は数値が大きいほど地域の剥奪度が高い、つまり社会的に不利な経済状況にある人々の割合が高いという指標である。これを各地域の人口で重み付けし、5 分位に分けた。第 1 分位が最も剥奪度が低く裕福な地域（高 SES 群）、第 5 分位が最も剥奪度が高い地域（低 SES 群）。がん罹患率は検診受診等の予防行動に影響を受ける。例えば、検診受診率の高い地域では、より多くの患者ががんと診断されるため、罹患率が高くなる。そのため、診断時の進行度を早期がん：上皮内＋限局、進行がん：領域＋遠隔転移に区分し、進行度別の年齢調整罹患率を ADI5 分位グループごとに算出した。部位・性別・診断時期別に、分散重み付き最小二乗法により ADI と進行度別罹患率の関連を分析し、ADI 第 5 分位地域と第 1 分位地域の罹患率差を推定した（Model 1）。

Model 1: $ASIR_i = \beta_{ADI} a_i + e_i$

また、診断時期により格差が拡大したか、縮小したかについて検討するために、診断時期と ADI との交互作用項で評価した（Model 2）。

Model 2:

$ASIR_{ij} = \beta_{ADI} a_i + \beta_{per} p_j + \beta_{ADIper} a_i p_j + e_{ij}$

（倫理面への配慮）

大阪府がん登録より入手した情報は匿名化された情報であるが、居住地住所等の情報を含むデータを扱う際は、スタンドアローン環境の PC において作業を行う。また社会経済指標等と突合後には個人同定が不可能な状態に変換する。

C．研究結果

性別・部位別・診断時期別・進行度別に ADI 第 5 分位（低 SES 群）と第 1 分位（高 SES 群）の年齢調整罹患率の差を分散重み付き最小二乗法における推定値を元に算出し図 1、2 に示した。進行度別年齢調整罹患率の分位ごとの傾向は性別・部位別に図 3～11 および表 1 に示した。前立腺がん以外の全ての部位で、進行がんの罹患率は ADI の高い地域（低 SES 群）ほど高かった。ADI 第 5 分位（低 SES 群）と第 1 分位（高 SES 群）間の進行がんの罹患率差は、男女とも 1993-1998 年診断の肺がんが最大で、それぞれ人口 10 万人対 12.0（95%信頼区間：9.4-14.5）、5.7（4.4-7.0）であった。しかし、その差は 1999-2004 年には統計的有意に減少し、他の部位の進行がん罹患率においても、統計的有意には拡大しなかった。一方、

早期がんの罹患率は、男性で前立腺、胃、大腸において、ADIの高い地域（低SES群）ほど罹患率が低かった。その傾向は前立腺で顕著であり、1999-2004年ではADI第5分位と第1分位間の罹患率差は−7.3と拡大した。女性では子宮頸がんにおいて、ADIの高い地域（低SES群）の罹患率が高かった。格差の縮小が最も大きかったのは、男性の胃・大腸の進行がんであったが、この傾向は女性ではみられなかった。

D．考察

大阪府がん登録資料より、前立腺以外の主要部位の進行がんの罹患率で社会経済格差が生じていることが明らかとなった。進行がんにおいては低SES群の罹患率は高SES群よりも高く、男性の早期がんでは、その逆の関連が見られた。

進行がんにおける社会経済格差の要因は、喫煙率やハイリスクな性行動などがん発症のリスクとなりうる行動の違いや、がん検診受診やがんの自覚症状への気づき（awareness）から医療機関へのアクセスの違いなど様々な要因が融合して影響していると考えられる。例えば、喫煙率に関しては平成26年度国民健康栄養調査において、男女とも世帯年間収入が600万円以上の群に比べ、200万円以上600万円未満、200万円未満の群の喫煙率が統計的有意に高かった。[1]喫煙をリスクとする肺がん罹患率の社会経済格差は喫煙率の違いにより部分的に説明可能であるといえる。

また、がん検診受診率についても、国民生活基礎調査をもとに、医療保険別にみた場合、共済組合加入者は市町村国民健康保険加入者よりも25〜40ポイント受診率が高い傾向にあった。[2]男性において、高SES群の方が低SES群より早期がん罹患率が高い理由としては、裕福な地域に住む住民の方が職場をはじめ、がん検診を受診する機会が多く、早期に診断された可能性が示唆された。しかし、この逆の相関関係は、女性では観測されなかった。女性においては、常勤として従事している人の割合が低いため、職場における検診受診体制の差の影響が男性ほど大きく出なかったと推察される。一方、進行がんの罹患率においては、女性ではどのがんにおいても低SES群の方が高SES群より高く、肺がん以外ではその格差は縮小していなかった。

米国の8つのがん登録データを用いた地域別進行乳がん罹患率は、マンモグラフィ施設密集度や高学歴者割合、英語識字率吐逆相関し、黒人割合と相関していた[3]。また、米国9つのがん登録データより、County-levelの剥奪指標4分位を用いて、詳細のステージ別罹患率および死亡率を分析した研究においては、高SES群の上皮内がん罹患率が低SES群よりも高く、その差は拡大傾向にあった。それ以外の進行がん罹患率や死亡率に関しては、格差が顕著ではなかった。[4]米国における子宮頸がんに関しては、死亡率・進行がん罹患率ともに格差が拡大傾向にある報告もある。[5]

がん登録資料より、診断されたがん患者内における進行度分布における進行がんをアウトカムにした研究においても、SESが進行がんでの診断に影響を与える結果が各種報告されている[6-8]。しかし、この研究デザインの場合、過剰診断による影響を大きく受けるため、人口を分母とした本研究のような進行度別罹患率による検討が望ましい。

がんにおける総合的なアウトカム指標は

死亡率である。がんにおける社会経済格差を検討する際には、がん進行度別罹患率だけでなく死亡率を最終アウトカムとし、他に生存率および喫煙率、検診受診率、医療アクセスなど各種関連指標と総合的にその関係性を分析しメカニズムを解明する必要がある。その上で、がんによる死亡の社会経済指標による格差を縮小するための効果的な手立てを検討できるといえよう。

現時点（H28年3月）で、利用可能なデータとして、本研究のように町字レベルまで使用可能であるのは、がん登録資料（生存率および医療アクセス）および近年の人口動態統計オンライン届出分のデータに限られる。国民生活基礎調査から得られる喫煙率およびがん検診受診率は都道府県レベル、自治体で行われるがん検診受診率は地域保健・健康増進事業報告の市町村レベルのデータに限定される。これらの指標に関しては、自治体におけるデータ提供体制の整備が急務であるが、サンプル調査に関しては空間的マイクロシミュレーションを用いた推定法を適用し、小地域における推定結果を用いる必要がある。

E．結論

大阪府がん登録資料より、進行度別がん罹患率の社会経済格差を分析した。高SES群に比べ低SES群では進行がんの罹患率が高かった。男性において、高SES群において早期がん罹患率が高い傾向も観測された。進行度別罹患率の社会経済格差は喫煙やハイリスクな性行動など、がん発症のリスクとなりうる行動の違いやがん検診の受診率の違いなどが影響していると考えられ、さらなる検討が必要であることが示唆された。

F．健康危険情報

なし

G．研究発表

1．論文発表

Ito Y, Nakaya T, Ioka A, Nakayama T, Tsukuma H, Uehara S, Sato KK, Endo G, Hayashi T: Investigation of Spatial Clustering of Biliary Tract Cancer Incidence in Osaka, Japan: Neighborhood Effect of a Printing Factory. J Epidemiol 2016, [in press].

Kinoshita F, Ito Y, Nakayama T: Trends in lung cancer incidence rates by histological type in 1975-2008: a population-based study in Osaka, Japan. J Epidemiol 2016:in press].

伊藤ゆり，中山富雄：肺がん生存率の国際比較．肺癌 2015, 55:266-272.

2．学会発表

伊藤ゆり，中谷友樹，近藤尚己，福井敬祐，中田佳世，井岡亜希子，宮代勲，中山富雄．大阪府におけるがん進行度別罹患率の社会経済格差：1993-2004年における格差の変化．第74回日本公衆衛生学会総会．2015:402 (P-0802-10)．長崎

Ito Y, Nakaya T, Kondo N, Fukui K, Nakaya K, Ioka A, Miyashiro I, Nakayama T, Rachet B. SOCIO-ECONOMIC DIFFERENCES IN STAGE-SPECIFIC CANCER INCIDENCE IN OSAKA, JAPAN: 1993-2004. 37th International Association of Cancer Registries, Annual Scientific Conference