厚生労働科学研究費補助金（がん対策推進総合研究事業（がん政策研究事業））
分担研究報告書

がん登録データに対する統計手法の開発に関する研究
（罹患報告の遅れ補正モデル）

研究分担者　加茂憲一　札幌医科大学医療人育成ｾﾝﾀｰ　准教授
研究協力者　福井敬祐　大阪医科大学研究支援ｾﾝﾀｰ　助教
研究協力者　伊森晋平　広島大学大学院理学研究科　助教

研究要旨

本研究においては，地域がん登録にがん罹患データが集約されるプロセスおよび、それが全国がん登録として集計されるプロセスにおいて必然的に発生するタイムラグに着目した。具体的には、MCIJ（Monitoring of Cancer Incidence in Japan）として一旦報告された値からの変更（追加・修正・削除）が発生するメカニズムに着目し、それが全罹患に対して占めるパーセンテージを考察対象とした。MCIJにおけるデータ収集では、最新の報告該当年における罹患のみならず、これまでに報告されている過去の分も同時に情報収集する。従って、過去に報告されていた数値に対して、その後に遅れ修正が発生してきた経緯を知ることができる。その傾向を統計学的手法によって分析することにより、罹患の最新報告から今後どの程度の遅れ報告による修正が発生するかを推測することが可能となる。分析にはANOVA（Analysis of Variance：分散分析）モデルを適用し、これまでのデータの経時的な傾向を表現する。そのモデルを将来の部分に延長することにより、遅れ発生を予測する手法を適用した。実際に愛知県における全がん男性の1993年から2015年の罹患について、MCIJ2003からMCIJ2015として収集したデータを用いて時系列の特徴を観察し、罹患報告の遅れを補正した。

A. 研究目的

日本におけるがん罹患情報は、都道府県規模の地域がん登録で収集された後に、全国がん登録として集約される。その際、データの収集や集約・照合作業の過程において必然的なタイムラグが発生する。その結果として、がん罹患数の公式発表後に、修正や追加が発生するケースが一定数存在する。例えば、全国がん登録における集約時までに登録が間に合わなかったケースは、罹患数発表後に追加される（追加登録）ことになる。あるいは、発表後にデータ入力に誤りが発見された場合は、全罹患数としての変化はないが、登録情報の修正が発生する（登録変更）。他にも、登録されていた情報に関して、新たに過去の罹患情報が判明した場合には、当該年の罹患から削除し、正式な年に登録しなおす必要がある（登録削除）。このように様々な事情に起因して、一旦締め切って集約した罹患情報に変更が発生するケ

ースが、一定割合存在することが知られている。このようなケースが発生した際には、その都度過去の報告値を書き直すというアプローチも考えられるが、後日書き換えられると分かっているデータに対する信頼性が低くなってしまうという問題点がある。

このような問題点に対して、本研究においては過去の罹患報告遅れの発生状況や特性を数理モデルにより表現することにより、今後の修正発生を予測することを試みた。具体的にはSEER（Surveillance, Epidemiology, and End Results）において同様の問題に対して適用されているANOVA（Analysis of Variance：分散分析）モデルを採用し、日本のデータに適用することにより、それを全国がん登録において将来発生すると考えられる報告遅れの予測を行った。

B. 研究方法

全国がん登録として公表されているがん罹患数は、地域がん登録の情報を基にしている。またMCIJ（Monitoring of Cancer Incidence in Japan）においては、報告する当該年のみならず、過去分の罹患数も収集されている。従って、報告年の罹患数と、将来の罹患数の両方の情報が得られる。これらの差を取ることにより、報告遅れに関する時系列の情報が得られる。これらの情報に対してANOVAモデルを適用することにより、罹患報告の遅れの部分に関する補正を試みた。

解析には、MCIJ2003からMCIJ2015において収集された、1993年から2015年のがん罹患に関する情報を用いた。例えば、罹患に関して最も初期の1993年罹患に関しては、MCIJ2003からMCIJ2015にかけて13回分の報告・修正データが存在するため、この13回分に関する時系列の特徴をANOVAモデルにより表現することになる。

j年に報告されるi年の罹患数を$I_{i,j}$と表す。ANOVAモデルにおける被説明変数としては、隔年において報告される同一年の罹患数の比

$$r_{i,(k+1/k)} = I_{i,k+1} / I_{i,k}$$

に対して$\log r_{i,(k+1/k)}$を採用する。例えば，2000年罹患に対して，2005年に報告される数と2006年に報告される数の比は$r_{2000,(6/5)} = I_{2000,2006} / I_{2000,2005}$となる。SEERの先行モデルにおいては、この比の特性に関して次の4種類の群が存在すると仮定している：

A：$r_{i,(3/2)}$　（3年／2年比）
B：$r_{i,(4/3)}$　（4年／3年比）
C：$r_{i,(5/4)}$　（5年／4年比）
D：$r_{i,(6/5)}$ 以降全て　（6年／5年比以降の全て）

ここで、A群からD群に対応するパラメータをβ_A，β_B，β_C，β_Dとして、隔年（k+1年とk年）の罹患数の比の対数を被説明変数とする次のANOVAモデルを設定する：

$$E[\log r_{i,((k+1)/k)}] = \beta_A x_A + \beta_B x_B + \beta_C x_C + \beta_D x_D$$

ここで、x_A，x_B，x_C，x_Dはそれぞれ、群A，B，C，Dに対応するダミー変数である。これら4変数には多重共線性が存在するため、切片項は設定しない。実データを用いて、このモデルに含まれる4つのパラメータを推定することにより、遅れが発生するメカニズムを表現する。最終的には、推定結果を用いて将来発生すると考えられる罹患の遅れ

報告が予測可能となる。

（倫理面への配慮）
本研究には倫理面への配慮を要する内容は含まれない。

C. 研究結果

MCIJ2003からMCIJ2015において収集された都道府県別のデータから、人口規模が大きく時系列としての情報も豊富な愛知県、その中でも男性の全がんに着目して解析を行った。まず、遅れて報告される罹患数にどのような特徴があるのかについて1993年罹患がどのように報告されてきたかを図１に示す。

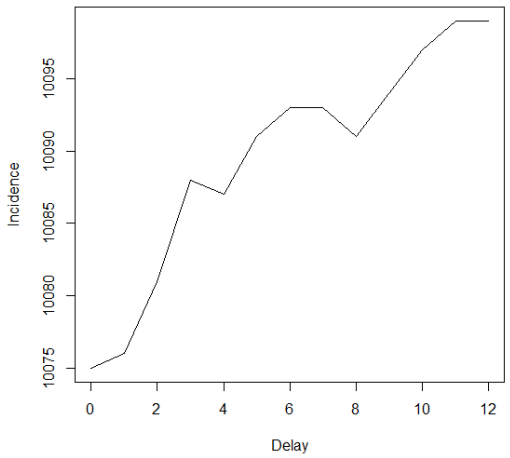

図１　1993年罹患の報告値

図１は縦軸を罹患数、横軸は遅れ年数（0をMCIJとして報告した年として規準化している）を表している。ここでは、MCIJ2003において報告された1993年罹患数を0年遅れと設定している。折れ線グラフは右上がりの傾向にあるため、遅れ報告は増加する（初期報告に積み上がる）という系統性が分かる。しかし、単調に積み上がる訳ではない年もあることも分かる（4年目と8年目のように前年から減少するとなる年も存在する）。一方で、罹患数全体に占める遅れ変動のパーセンテージに着目すると、10,000人の規模に対して20人程度の変動（0.2パーセント程度）であるため、MCIJ2003以降において劇的に罹患数が変動する訳ではないことも分かる。

前述のANOVAモデルを適用するにあたって、報告される罹患数における隔年比（1年間隔の比）の特徴を図２に示す。

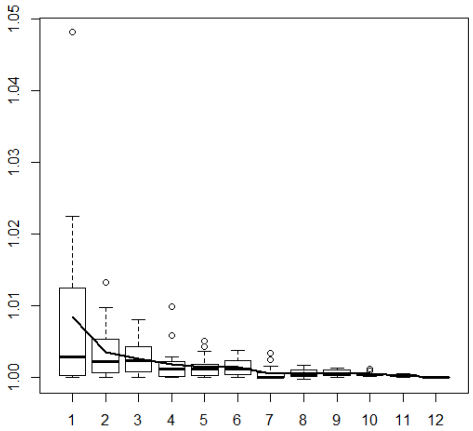

図２　罹患数の隔年比

図２において、縦軸は隔年の罹患数の比（k+1年罹患 ÷ k年罹患）、横軸は遅れ年を表している。例えば、一番左側のボックスプロットは初期報告と1年後の報告値の比を、二番目のボックスプロットは1年後と2年後の報告値の比を表している。折れ線は平均の挙動を表している。基本的に比は1以上であり、これは遅れ報告が前年にプラスされる傾向が強いことを意味しており、図１で観察した傾向が、1993年罹患以外においても

継続されていることを意味する。一方で、遅れ年が増えるに従って、罹患数の比は1に近づくことも分かる。これは年が経つに従って、遅れ報告が発生しなくなってくる傾向にあることを意味している。これは図１における右上がりの傾向が、後年になるに従って緩やかになることに対応している。同時にボックスプロットの幅も、遅れ年が増えるに従って狭くなる傾向があり、これは分散が後年になるに従って小さくなることを意味している。

次にANOVAモデルに基づいて未知パラメータを推定すると

$$\beta_A = 0.0083,$$
$$\beta_B = 0.0035,$$
$$\beta_C = 0.0026,$$
$$\beta_D = 0.0008$$

という結果が得られた。全てのパラメータが正値であることから、報告年以降の補正分はプラスとして現れる（追加される）傾向にあると言える。また、パラメータ間の大小関係に着目すると

$$\beta_A > \beta_B > \beta_C > \beta_D$$

であることから、遅れ修正（追加）が発生する頻度が高いのは、報告年直後であり、年が経つに従って最終的な数価に漸近することが分かる。これは図２における観察結果と一致する。

最後に、推定されたパラメータを用いて罹患数の予測を行った結果を図３に示す。

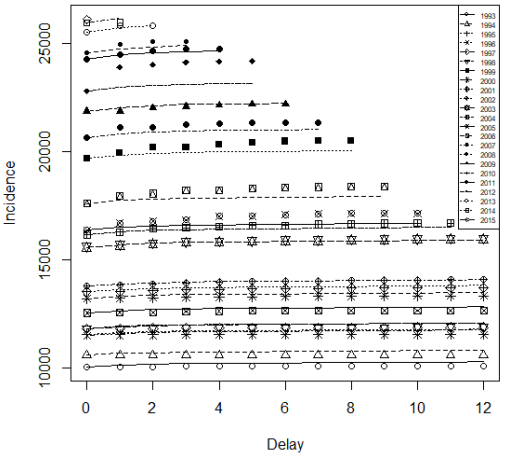

図３　罹患数の実測値とANOVAによる予測値

図３の縦軸は罹患数、横軸は遅れ年数（0をMCIJとして報告した年で規準化している）を表している点は図１と同様である。実測値をプロットで、ANOVAモデルによる予測値を折れ線で表している。例えば最も下側の「○」プロットは1993年罹患についてMCIJ2003からMCIJ2015にかけての報告値（図１と同じ）を表し、実線は同年罹患についてのANOVAモデルによる予測値を表している。モデルの特性上、遅れが0年（MCIJへの最初の報告年）においては実測と予測が一致し、その後から両者のズレが発生する。

愛知県における男性の全がん罹患数は経年的に増加傾向にあるが、それぞれの罹患年について、遅れ補正により若干の上積みが発生していること、そしてモデルによってその傾向がある程度再現できていることが分かる。

D. 考察

本研究では、ANOVAモデルを用いて、がん罹患数の事後修正パターンを再現することを試みた。実データ解析としては、2003年から2015年の罹患について、愛知県の男性における全がんのデータを用いた。その結果を示す図３では、概ねの特徴は再現されていると考えられるが、現実と乖離している箇所も存在する。その原因として現時点で考えられる問題点および将来的な発展について記述する。

まず、モデルについて考察する。今回SEERで採用されているANOVAモデルを採用した。被説明変数を隔年の罹患数の比の対数、説明変数を遅れ年数に対応した4つのダミー変数とした。説明変数の4つのダミー変数についてはSEERの流儀をそのまま用いたが、この分類は作為的であり、日本の状況を踏まえた上で再考を要する箇所であると考えられる。具体的には、図２のボックスプロットから、どのような遅れ年に関するグルーピングを行うのが最適かを検証する必要がある。また、最終的には罹患数の隔年比が1に収束することから、予測に何年遅れまでを用いるべきなのかについても考察する必要がある。また、本モデルは報告遅れ年のみを説明変数としているため、初期報告が決まるとその後の変動は比例的に決定される。つまり、初期報告が多い年は、その後も一定の比率で遅れ分が積み上がり、他の年との逆転が起きることはない。言い換えると、初期報告の大小関係のみが、その後の傾向を全て決定する（図３のように層状の折れ線グラフになる）という特徴がある。これはモデルとして硬すぎる性質であるため、改良の余地が残されている点であると考えられる。

次に実データの活用について考察する。今回は愛知県の男性について、年齢階級を考察しなかったが、MCIJにおいては年齢階級別でデータが提出されている。年齢に依存して遅れ発生のメカニズムが変化する可能栄としては、例えば若年層における予後の良さが遅れの多発を招いているような特性があるならば、年齢に関する説明変数を導入することによる推定の改良が可能になる。また、本解析には罹患年に関する変数も含まれていない。罹患年に関する要素は、例えばランダム効果モデルなどにより導入することが可能になると考えられる。罹患報告年に着目すると、2003年罹患については2008年3月の報告と約5年遅れであったものが、2015年罹患については2018年9月の報告で約3年半の遅れに変化する。これはデータ集約プロセスの改良や、タイムリーな罹患情報のニーズに応えるために、報告までのラグを小さくしてきた成果である。従って、本研究で用いたANOVAモデルの被説明変数である「隔年の罹患数の比」について、「隔年」の間隔が一様でない。例えば、2003年罹患が2008年3月報告であり、2004年罹患が2008年12月報告であるので、ここでの「隔年」は9か月である。一方で、2005年罹患が2009年9月報告であり、2006年罹患が20010年9月報告であるので、ここでの「隔年」は1年である。この2つの例を比較すると「隔年」に3か月のずれが発生している。このように「隔年」の期間の不均一性が推定に悪影響を与えている可能性は否定できない。今後は、隔年の取り扱いについても、モデルまたは変数の設定法に関する改良が必要である。

E. 結論

本研究では、がん罹患数について、タイムリーな報告後に発生する修正パターンに着目し、ANOVAモデルを用いてそのパターンを再現することを試みた。2003年から2015年の罹患について、愛知県の男性における全がんのデータを用いた実解析の結果を図３に示す。フィッティングは概ね良いが、現実と乖離している箇所も存在する。その原因としては、モデルやデータの利用法が未だ洗練されておらず、データの有する特性をフルに活用できていないからであると考えられる。具体的な内容は「D. 考察」で言及した通りであるが、本研究テーマは未だ新しい分野でもあり、今後の発展が期待される。

SEERの先行研究において本研究の意義として挙げられているのは、罹患の短期予測である。タイムリーな罹患数を報告するために、数理モデルを用いた試みがなされてきたが、短期予測では「長期トレンドを再現する形での延長」に加えて「直近の突然な変化に敏感であること」が重要な要素となる。この点に関して、罹患の遅れが補正されていないデータを用いた場合、まさに直近のトレンドに変化（過小評価）が発生しやすくなる。予測において折れ線回帰（joinpoint）を施す場合には、直近年に近い部分に不要な節点が発生しやすくなり、それが短期予測の結果を大きく狂わせる原因となりかねない。このような問題点を解決するためにも、罹患の遅れ補正は重要な研究テーマである。

F. 健康危険情報

（分担研究報告書には記入せずに、総括研究報告書にまとめて記入）

特になし

G. 研究発表

1. 論文発表

R.Tanabe, K.Kamo, K.Fukui, S.Imori. Statistical inference for estimating the incidence of cancer at the prefectural level in Japan. Jpn J Clin Oncol, 49 (5), 481-485, 2019.

2. 学会発表

（発表誌名巻号・頁・発行年等も記入）

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