厚生労働科学研究費補助金（第３次対がん総合戦略研究事業）
分担研究報告書

がん罹患の動向分析

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
研究分担者　片野田耕太　国立がん研究センターがん対策情報センターがん統計研究部　室長
研究分担者　雜賀公美子　国立がん研究センターがん予防・検診研究センター検診研究部　研究員

研究要旨
がん罹患の挙動に影響を与える時間に関連する要因の統計解析を行う。昨年度報告書における罹患リスク視覚化モデルの発展形として、５年の短期予測に着目した。日本におけるがん罹患数の報告には約５年の遅れがあり、これを統計モデルによって補填し、タイムリーな数値を報告することが目的である。ポアソン回帰モデルにおける変数選択においては、従来から用いられているAICを改良し、予測に特化した新たな規準量（PAIC）を算出した。そして仮想的な状態を３種類設定し、予測結果の比較検討を行った。男性の肝臓がんに関する解析から、AICよりもPAICの方が実測と予測のずれが小さいことが分かった。

A．研究目的

がんの挙動には様々な要因が影響を与えている。特に時間に関する要因に着目すると、年齢・時代・出生コホートの３要因が知られている。これらを取り扱う時系列分析は、過去の特性を基にして今後の動向を把握するために必要不可欠な手段である。解析目的は時系列の特性を表現することにあるが、手法が複雑になれば自動的に特性の表現（アウトプット）も複雑になる。昨年度の報告書においては、がんの時間に依存する特性をシンプルに表現する手法として、年齢と時代を基底とするリスク曲面を用いるものを提案した。本報告書においては、この応用として罹患数の短期予測を試みる。

日本におけるがん罹患数の報告には約５年の遅れがある。これを補填しタイムリーな情報を得るために、短期の予測が試みられている。短期予測は長期予測（将来予測など）の一部分ともみなせるが、短期に特化した特性や手法が存在するため、それらを用いるのが一般的である。例えばKatanoda et al(2014)“Short-term projection of cancer incidence in Japan using an age-period interacti on model with spline smoothing”においては年齢と時代の交互作用を持つsplineを用いた手法が日本のデータに対して適合するものとして提案されている。

本報告書においては、昨年度の報告書におけるリスク視覚化モデルの応用として短期予測を試みる。用いる基本的な統計モデルは、Katanoda et al (2014)と同じく人口をオフセットしたポアソン回帰である。回帰モデルにおける変数選択においては、赤池情報量規準（AIC）が広く用いられるが、具体的な予測年数が判っている場合（今回は５年）に特化した形のAIC（Predictive AIC:PAIC）を新たに算出し、どのように予

測結果が変わるかを比較検討した。

B．研究方法

昨年度の報告書において報告したがんリスクの視覚化モデルを用い、その外挿により短期の予測を行う。これを地理的な概念で捉えると地図の外挿であることを考えると、あまり広いエリア（長期）の予測に適用することは不適切と考えられる。今回の5年程度の短期予測が限界であろう。

カレンダー年 p において年齢 a の罹患数と人口をそれぞれ y_{ap}，z_{ap} とする。罹患数 y_{ap} が z_{ap} をオフセットとするポアソン分布に従うと仮定すると、パラメータ λ_{ap} を用いて

$$y_{ap} \sim \mathrm{Poisson}(\lambda_{ap}\ z_{ap})$$

となる。対数線形性を仮定すると

$$\log\lambda_{ap} = \boldsymbol{\beta}'\mathbf{x}(a,p)$$

となる。ただし $\boldsymbol{\beta}$ は未知パラメータ、$\mathbf{x}$ は a と p からなる説明変数ベクトルである。例えば1次の交互作用を設定する場合には

$$\boldsymbol{\beta}'\mathbf{x}(a,p) = \beta_0+\beta_1 a+\beta_2 p+\beta_3 ap$$

となる。今回の解析においては最大4次の交互作用まで含むモデルをフルモデルと設定し、変数選択を行った。このようなポアソン回帰モデルにおいて変数選択の際に用いられるのが次の AIC である。

$$\mathrm{AIC} = -2\log L(D;\boldsymbol{\beta})+2k$$

ここで L は対数尤度、D は変数（説明変数と被説明変数）、k は設定したモデルに含まれる未知パラメータの個数を表す。例えば、1次の交互作用モデルにおける未知パラメータは β_0，β_1，β_2，β_3 の4つであるので k=4 である。上記の AIC は実測のデータのみを用いたモデル選択規準量であるが、今回のように予測が解析目的である場合には、AIC を予測に特化した形に修正した次の PAIC の方がより良いパフォーマンスが期待できる：

$$\mathrm{PAIC} = -2\log L(D;\boldsymbol{\beta})+k+\mathrm{tr}(\Gamma_{\mathbf{W}}\Gamma_{\mathbf{X}}^{-1})$$

ここで $\mathbf{X}$ は実測の範囲の説明変数、$\mathbf{W}$ は予測部分も含む説明変数、$\mathbf{V}_{\mathbf{X}}=\mathrm{diag}(\mathrm{var}[y_1],\ldots,\mathrm{var}[y_n])$，$\Gamma_{\mathbf{X}}=n^{-1}\mathbf{X}'\mathbf{V}_{\mathbf{X}}\mathbf{X}$（ただし n は観測数）であり、$\Gamma_{\mathbf{W}}$ は変数 $\mathbf{W}$ に対して $\Gamma_{\mathbf{X}}$ と同様に定義したものである。

解析においては1975～2007年の男性年齢階級別の人口と肝臓がん罹患数のデータを用いた。ただし、若年および高齢における不安定さを除くために、40～84歳データに限定した解析を行った。

C．研究結果

年齢と時代およびこれらの交互作用項を含むモデルを用い、出生コホート効果の強い肝臓がん（男性）に関する5年予測を行った。出生コホート項を含まないモデルにおいても、出生コホート効果の強い肝臓がんの特性が表現できるかがポイントである。

まずは、出生コホート効果の存在および強さを、前年度報告書における手法を用いリスク曲面として表現して確認する（図1）。左側が地理的加重一般化線形モデルによるもの、右側がパラメトリックモデルによるものである。

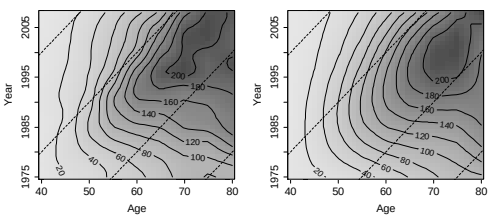


図1．肝臓がん（男性）のリスク曲面

図 1 は、横軸が年齢、縦軸が時代を表し、その基底の上にリスクの高低を曲面として表現したものである。リスクの高低を色の濃淡と等高線で表現している。濃色の部分が高リスクであることを意味する。また等高線上の数値は 10 万対の人数である。左下から右上にかけての破線は同一出生コホートを表し、右下から 20 年間隔に 1900 年、1920 年、1940 年、1960 年出生コホートである。多くの先行研究で指摘されている通り、昭和 1 桁生れ世代における高リスク効果が確認できる。モデルには出生コホート効果に対応する直接的な項は含まれていないが、交互作用項が代替の役目を果たしていると考えられる。

次にこのモデルを用いて 5 年予測を行った。ただ、2007 年からの 5 年予測（2012 年予測）を行っても結果の妥当性を検証できない。なぜなら実測のデータが存在しないからである。そこで仮想的に次の 3 パターンを用意し、実測との「ずれ」を計測した：

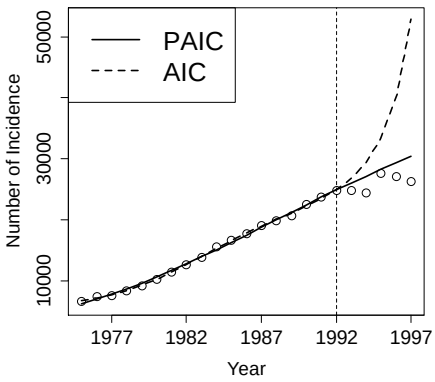
① 1975～1992 年 ⇒ 1997 年予測
② 1975～1997 年 ⇒ 2002 年予測
③ 1975～2002 年 ⇒ 2007 年予測

各設定において選択されたモデルは、

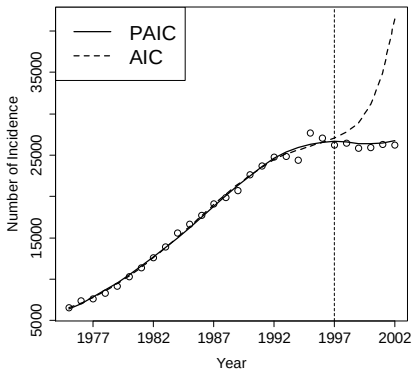
① AIC：年齢 4 次、時代 4 次
　PAIC：年齢 4 次、時代 2 次
② AIC：年齢 4 次、時代 4 次
　PAIC：年齢 4 次、時代 3 次
③ AIC・PAIC：年齢 4 次、時代 4 次

であった。次に①～③についての予測結果を図 2 に表す。

①



②



③

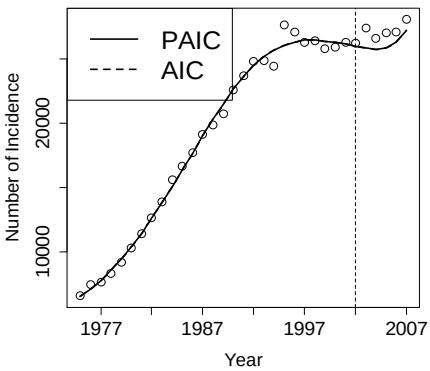


図 2．3 パターンの予測結果

横軸がカレンダー年、縦軸が罹患数を表す。また、プロットが実測値、破線が AIC より選択されたモデルによる予測値、実線が PAIC により選択されたモデルによる予

測値である。①と②に関しては AIC と PAIC による結果が異なり、③に関しては一致した。結果の異なった①と②に着目すると、予測の精度を実測との差異で測るとすればPAIC の方が優れた結果である。実際にはPAIC の方がよりシンプルなモデルを選択しており、AIC の特性である複雑なモデルを選びやすいという点が修正されていることが伺える。①の予測結果をリスク曲面で表すと図 3 となる。

(a) (b) (c)

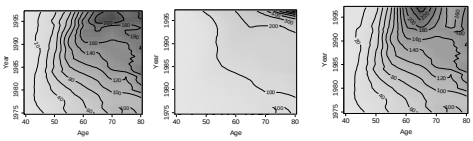


図 3　予測結果のリスク曲面①

(a)は実測、(b)は AIC による予測、(c)は PAIC による予測をリスク曲面で表現したものである。AIC による結果は、出生コホート効果が再現されていない。また、複雑なモデルが選択されていることに起因して、エッジ（1997 年近傍）の部分に極端な挙動が表れている。これらが予測の過大評価につながっていると考えられる。次に②の予測結果をリスク曲面で表したのが図 4 である（(a)-(c)の意味は図 3 と同じ）。

(a) (b) (c)

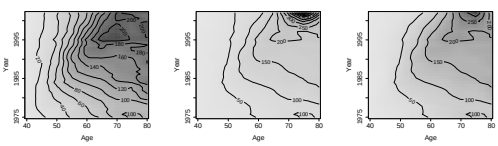


図 4　予測結果のリスク曲面②

①の場合と同様に、AIC による予測はエッジの部分が柔らか過ぎる挙動を示す一方で、PAIC による結果は安定している。

D. 考察

解析に用いたモデルは一般化線形モデル（ポアソン回帰モデル）であり、AIC によりモデル選択が行われることが多い。しかし AIC は実測のデータに対するモデルのあてはまりから最適なモデルを選択するための規準量であり、今回のように外挿が目的である場合には、その目的の達成に特化した選択規準を用いることにより予測パフォーマンスの向上が期待される。今回の外挿は 5 年先と具体的な予測年数が決まっていることから、前述の PAIC が適用可能となる。

今回仮想的に 3 パターンを用意し、AIC および PAIC によるモデル選択を行い、それらに基づく予測結果と実測値の乖離を観察した。その結果、2 パターン（①と②）において AIC と PAIC で異なる結果を得た。実測値により近い予測値が得られたのは共に PAIC であり、モデルとしてはシンプルなものを選択していた。一方で③のみ AIC による結果と PAIC による結果が一致した。その理由として、③はもっとも長期間の実測データを用いており、予測部分のパーセンテージが低いことが挙げられる。PAIC は予測部分を考慮した規準であり、そうでない場合（$W=X$ の場合）には $\mathrm{tr}(\Gamma_W\Gamma_X^{-1})=k$ となり AIC に一致する。③はこれに近い状態であるため、AIC と PAIC の間に相違が発生しなかったと考えられる。

現在、短期予測に関しては前出の Katanoda et al (2014)による手法が日本のデータに良く適合することが知られている。短期予測に関しては絶対的な手法が存在しないため、様々なモデルを比較検討しながら最適な手法を模索する必要があると考えられる。

E. 結論

現在、日本におけるがん罹患の報告は5年遅れであり、この即時性の問題を解決する手法の1つに短期予測がある。このような試みはAmerican Cancer Society (ACS)でも行われているが、短期予測において用いる統計手法としては確たるものが存在しないのが現状である。実際にACSで用いられている統計手法も何度か変更されてきた。日本においてはKatanoda et al (2014)においてspline交互作用モデルが適合すると報告されているが、今後も更なる手法の改良および他モデルの通用可能性について議論を深めていく必要がある。本報告書では、その1つの候補として「予測年数が確定している」という限定された状況において、従来のAICを改良したバージョンであるPAICを提案し､実際にデータ解析を行った。男性の肝臓がんで解析した結果、AICよりは優れた結果が得られた。今後の課題としては他の部位に関する解析も継続すること、および他のモデルとの比較検討を行うことが挙げられる。今回は行えなかったspline交互作用モデルとの比較も今後は必要となるであろう。

F. 健康危険情報

（総括研究報告書にまとめて記入）

G. 研究発表

1. 論文発表

1) K.Kamo, H.Yanagihara, K.Satoh, Bias corrected AIC for selecting variables in Poisson regression models, Communications in Statistics, 42, 1911-1921, 2013.

2) K.Katanoda, K.Kamo, K.Saika, T.Matsuda, A.Shibata, A.Matsuda, Y.Nishino, M.Hattori, M.Soda,A.Ioka,T.Sobue,H.Nishimoto, Short-term projection of cancer incidence in Japan using an age-period interaction model with spline smoothing, Japanese Journal of Clinical Oncology, 44 (1), 36-41, 2014.

2. 学会発表

1) 加茂憲一, 佐藤健一, 冨田哲治, 伊森晋平, がんリスクの予測を目的とした変量選択の試み, 統計関連学会連合大会, 大阪, 2013.

2) 雑賀公美子, 松田智大, 松田彩子, 斎藤博, 子宮頸がん罹患率の時系列解析, 地域がん登録全国協議会　第22回学術集会, 秋田, 2013

3) 雑賀公美子, 西本　寛, 松田智大, 斎藤博, 地域がん登録における検診由来がんの特徴, 第36回日本がん疫学・分子疫学研究会総会, 岐阜, 2013.

H. 知的財産権の出願・登録状況

1. 特許取得　なし

2. 実用新案登録　なし

3. その他　なし