

がん登録データに対する統計手法の開発に関する研究

加茂 憲一 札幌医科大学 医療人育成センター 准教授

研究要旨

本研究では、全国がん登録において収集されるデータを用いた統計解析に着目した。R2年度の研究報告においては、特にがん登録に関連するデータにおける時系列の挙動を表現するための統計モデルに着目した。具体的には「がん登録における登録率」と「生涯リスク」の2つのテーマに着目し、これらの時系列の特性に関する統計解析を行った。地域がん登録における登録率に関しては、罹患登録の完全性を回帰モデルによって推定する手法を用いて、2008～2015年における都道府県規模の、罹患・死亡・DCN数に基づく回帰モデルにより、地域がん登録における登録率を推計し、その経年変動を観察した。生涯リスクについては、生命表による年齢累積の罹患・死亡リスクを推定する手法を1975～2015年の全国罹患数・死亡数に適用した。それらの経年変動を年齢とカレンダー一年を座標とするHeatmapで描くことにより表現した。

A. 研究目的

本研究は、がん登録により収集されたデータに対して、統計解析による二次利用を行うことを目的とした。特に経年変動に関する統計解析に着目し、2つの研究テーマ

1. がん登録における登録率推計の経年変動
2. がん累積リスクの経年変動

に着目した解析結果を紹介する。

テーマ1については、登録の完全性に関する指標であるMI比（死亡／罹患比）とDCN割合（DCN数／罹患数）の関連性を定式化したものに対して回帰モデルを適用し、DCN割合が0%におけるMI比を表す未知パラメータを推定する。ここでのMI比とDCN割合の関連性については、登録の有無と生死に関する $2 \times 2 = 4$ 群における関係性を整理することにより、MI比とDCN割合の2変数と、未知パラメータという形で再構築する。このことにより、登録の完全性に関する2指標（DCN割合とMI比）を線形・非線形の関連性で結びつけることが可能となる。ここからDCN割合が0%でのMI比に対応するパラメータを推定し、死亡登録は人口動態統計により完全であることを用いて、登録が完全な状態での罹患数が推定可能となる。この罹患数に対して、報告された罹患数のパーセンテージを算出することにより、登録率が推定される。経年的にはDCN割合が低減してきており、このことは登録の完全性が上がってきていることを示唆している。この点について、上記の統計モデルを用いた経年観測を行うことによる確認を行った。

テーマ2については、がん罹患や死亡に関して、生命表を用いた年齢累積リスクや生涯リスクを推定する手法が確立されている。この手法を経年的に適用することにより、累積リスクの（年齢、時代）に関する変動を観察することを目標とした。加齢に伴うリスク上昇は通常折れ線グラフで表現されるが、それを経年的に重ねると表現が煩雑になるため、Heatmapを用いた表現を試みる。ここでHeatmapとは、変数の組合せを仮想的な番地と見做し、その番地における高低を色合いや等高線を用いて地図のように表

現したものである。我々は日常から地図を見慣れていることから、Heatmapによる表現から様々な特徴を容易に見出すことが可能である。また、生命表では時代に伴った年齢分布についても調整されることから、年齢調整されたがんリスクの挙動を観察するという意味合いも有する。

B. 研究方法

テーマ1については、登録の完全性に関する指標として用いられてきたMI比とDCN割合に、登録の有無（2群）と生死（2群）に関する4パターン（2群×2群）を集約する。このとき、4つの変数を2変数に集約することになるが、理論上は完全な集約は変数が不足するため不可能である。そこで、集約できなかった部分は未知パラメータに組み込むことにより推定する。未知パラメータの一つを「DCN割合が0%におけるMI比」と設定し、変数（DCN割合とMI比）の関連性を線形関数で定式化する。一方で、別のパラメータとして、登録・未登録群でのがん死亡率の差異を加えた形式で非線形モデルを構築する。どちらのモデルも、死亡が罹患からある一定の確率で生起する二項分布を想定している。パラメータ推定においては、罹患数の規模で都道府県の重み付けした最尤推定を行う。二項分布を仮定するため、回帰構造としては線形と非線形2種類のロジスティック回帰モデルを用いる。線形モデルと非線形モデルの優劣については、モデル選択で通常用いられるAICによる判断は、パラメータ数のみでモデルの複雑さが評価できないケースであるため適用できない。そこで、ジャックナイフ推定に基づくクロスヴァリデーション規準を用いたモデル選択を行う。

以上のプロセスを、2008～2015年の都道府県別罹患数、死亡数、DCN数データに適用し、DCN割合が0%におけるMI比を推定した。その結果と、人口動態統計の死亡数（死亡統計の完全性は100%であるという前提で）を用いることにより、真の（完全登録での）罹患数を推定し、報告された罹患数と比することにより登録率を推定した。

テーマ2については、年齢累積の罹患・死亡リスクに着目する。0歳人口から、加齢に伴った死亡数を、癌死亡と他死因死亡を分割して減ずることにより、累積死亡リスクが推定される。一方で、罹患に関しては死亡に関する生命表情報を基に、人口をCancer-freeに置き換えた生命表を再構築することにより罹患の累積リスクを推定する。カレンダー一年を固定することにより、年齢依存の累積リスクが推定されるが、その経年変動の表現についてはHeatmapを用いることが適切と考えた。Heatmapは変数の挙動を(x軸、y軸)の座標を固定した上に、色や等高線を用いて地図のように表現したものである。地図は身近に用いられている表現法であり、全体的な挙動に加えて細かな特異的な特徴を捉えやすいという特徴がある。ここでは、x軸をカレンダー一年、y軸を年齢と設定し、累積リスクを仮想的な標高として表現した。

C. 研究結果

テーマ1について、2008年、2010年、2012-2015年に関する推定結果を図1に示す。図1は、x軸がDCN割合、y軸がMI比を表している。メインタイトルに該当年と解析に用いた都道府県数を表す(例えば左上の2008年の結果は34都府県のデータを用いた)。2012年以降は全47都道府県データを用いた。プロットは各都道府県の実測値を表し、直線が線形モデル、曲線が非線形モデルを表す。また、それぞれのモデルにより推定された「完全登録でのMI比」を、破線で表す。線形・非線形モデル共に座標(1,1)を通り、切片が「完全登録でのMI比」を意味する。基本的に線形モデルは非線形モデルに比べて低い「完全登録でのMI比」(=低い登録率)が推定される傾向にある。それは、どちらのモデルも(DCN割合=100%, MI比=1)が不動点であることと、非線形モデルにより推定される曲線が下に凸だからである。

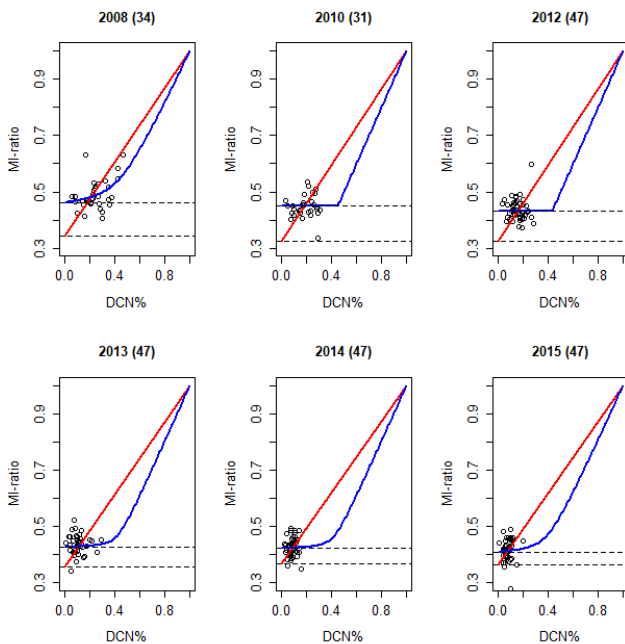


図1 MI比-DCN割合プロット

図1により推定された「完全登録でのMI比」に同年の全国がん死亡数を乗じることにより、登録完全

での全国がん罹患数を算出した。次に、推定された罹患数と、MCIJとして報告されている全国がん罹患数の比により、地域がん登録の登録率を年別(2008-2015年)に推定した。その結果を図2に示す。x軸がカレンダー一年、y軸が登録率を表し、2本の折れ線がそれぞれ線形モデルと非線形モデルにより推定された登録率を表す。線形モデルでは、2010年まで75%前後であった登録率が、それ以降上昇傾向となり、最終的な2015年には約90%と改善された。一方で非線形モデルによる登録率は100%を超える値が推定された。これは非現実的な推定値であり、線形モデルを適用すべきであることが示唆された。その原因については後に記述する。

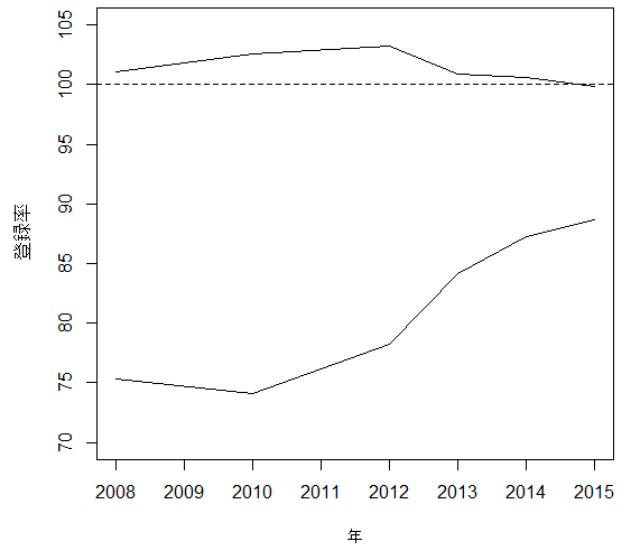


図2 登録率の経年変動

次にテーマ2に関する結果を示す。1975-2015年の人口、全死亡数、がん死亡数、がん罹患数を用いてカレンダー一年を固定して、性別に累積死亡・罹患リスクを推定した。使用データにおける最新2015年に関する結果を図3に示す。横軸は年齢を、縦軸は累積リスクを表す。横軸の84歳以上の箇所は生涯リスクを表す。4本の折れ線はそれぞれ性別の罹患、死亡リスクを表す。

関してはやや右下がりであることが分かる。

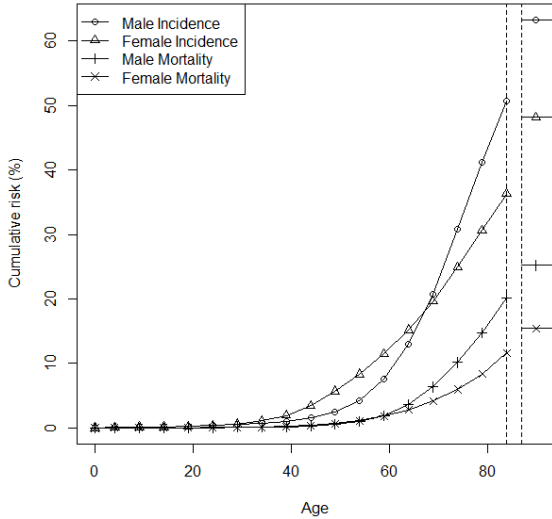


図3 累積罹患・死亡リスク

次に、横軸の最右部分である生涯リスクについての経年変動を表したのが図4である。横軸はカレンダー一年を表し、縦軸は生涯リスクを表す。4本の折れ線は、性別の生涯罹患、死亡リスクを表す。

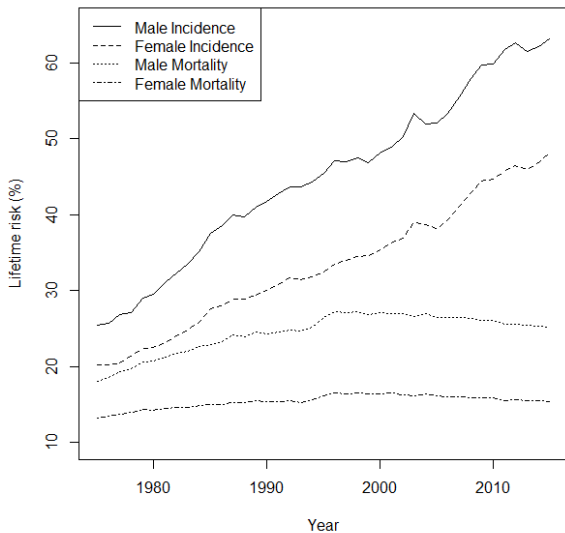


図4 生涯リスクの経年変動

このように、到達年齢を特定した経年変動は折れ線グラフにより表現できるが、年齢別の折れ線グラフを重ね描きするとグラフが煩雑となり特性を把握することが困難となる。そこで2つの時間に関する変数（年齢とカレンダー年）を仮想的な座標としたHeatmapを用いて累積リスクを表現することを試みる。図5は男性の罹患・死亡に関する累積リスクを表している。左側が死亡、右側が罹患でありどちらも横軸がカレンダー年、縦軸が年齢を指し、累積リスクの高低を色と等高線により表現した。色の濃度が濃いエリアが高リスクである。また、等高線上の数字の単位は%である。どちらも等高線は水平方向に走っているが、死亡に関してはやや右上がり、罹患に

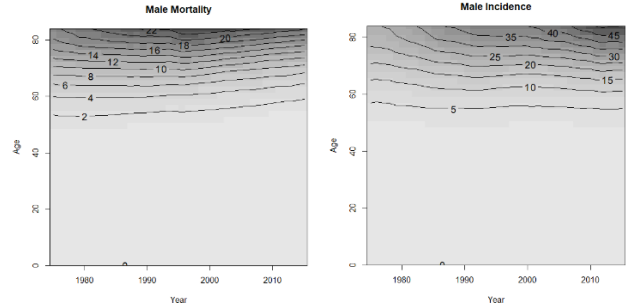


図5 男性の累積罹患・死亡リスク

次に、図6は女性の罹患・死亡に関する累積リスクを表したものである。全体傾向は男性と類似しているが、罹患の右下がり傾向が男性に比べて大きいことが観察される。

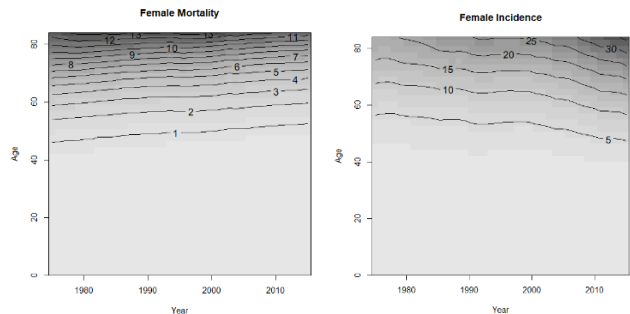


図6 女性の累積罹患・死亡リスク

以上は累積リスクに関する結果であるが、罹患に関する生命表を作成するプロセスでは、Cancer-free人口も生成される。全人口からCancer-free人口を減じることにより、がん生存（がんに罹患しかつ生存）人口が算出できる。図7はがん生存人口が全人口に対して占めるパーセンテージに関するHeatmapであり、等高線上の数値の単位は%である。左側が男性、右側が女性である。

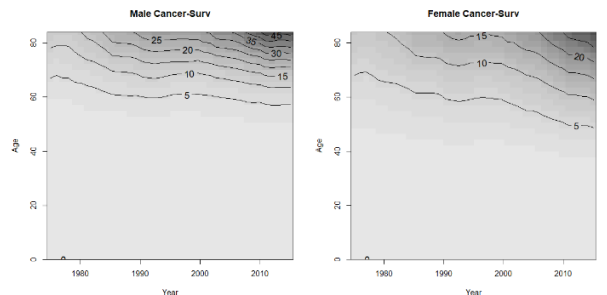


図7 がん生存割合のHeatmap

がん生存者割合のHeatmapの等高線は男女ともに右下がりの傾向が観察された。この右下がり傾向は女性の方に強く観察された。

D. 考察

まずテーマ1について考察する。推定された登録率は、線形モデルの下で2008年に75.4%であったも

のが、2015年に88.7%へと改善された（図2）。このことは、図1においてDCN割合が経年的に減少していることから類推されるが、ロジスティック回帰モデルを用いて登録率を推定することにより登録率が改善されていることが定量化・可視化できた。その一方で、非線形モデルを用いると100%を超えるという非現実的な登録率が推定された。このことは、実測のDCN割合が0%に近づいたために、0近傍を水平に近付けるような極端な回帰係数が推定されたことに起因していると考えられる。例えば、2008年の推定においては、完全登録でのMI比を表すパラメータは2.44と現実的なものである一方で、登録・未登録群間の死亡率の比に対応するパラメータは0.005と極端に小さい。このパラメータは、本来1に近いのが現実的である（線形モデルは、このパラメータを1に固定したものである）。実測値がDCN割合0%の近傍に集まった上に、その限られたエリアで水平方向のバラツキが発生したため定義域 $[0,1]$ における全体的な推定でなく、（DCN割合=100%、MI比=1）を固定した状態での0近傍におけるローカルな推定に強い重みが付いたことが原因であると考えられる。

次に、テーマ2については年齢に関する累積罹患・死亡リスクをターゲットとした。生命表を用いた本手法は、本来観測年を固定した上で年齢に伴ったリスク変化を観察するものであるが、その経年的な変動を観察するためにHeatmapを用いた。このことにより、一定のリスクに到達する年齢については、罹患では下がる傾向、死亡では上がる傾向にあることが判明した。このことは、累積リスクの経年変動に関して、罹患と死亡の間に乖離が現れてきたことを意味するが、それはがん生存割合（人口とCancer-freeの差が全人口に占める割合）に関するHeatmapからも確認できる。具体的には一定の死亡リスクに到達する年齢は経年的に上がる一方で、罹患に関しては下がる結果として、経年的に罹患と死亡の間に乖離が発生し、このことはがん生存の割合を押し上げていることに対応していると考えられる。性別に傾向が異なる点に関しては、特に片性部位のトレンド（乳がんや子宮がん）に関して近年の若年化が進んでいることを反映していると考えられる。経年傾向

を折れ線グラフにより表現することのみでは観察し難いこれらの特性が、Heatmapにより可視化することができた。

E. 結論

本研究は、がん登録データに対して統計モデルを適用することを目的としている。本報告は特に時系列の挙動を観察することに主眼を置き「登録率の経年変動」と「累積リスクの経年変動」に関する結果を報告した。登録率は改善傾向にあり、約10年で75%から90%へと改善がみられた。改善に関する傾きも頭打ちになっておらず、最新の登録においては登録率が90%と完全登録に極めて近い状態になっていることが予想される結果であった。

累積リスクの経年変動については、罹患年齢が下降傾向、死亡年齢が上昇傾向であり、その乖離部分である「がん生存」のパーセンテージが経年的に増加傾向にあることが観察された。罹患年齢の下降傾向は検診による早期発見による診断年齢の若年化の影響、死亡年齢の上昇傾向は治療の向上による予後の改善効果を反映していると考えられる。年齢とカレンダー一年という二種類の時間的要素について、Heatmapを用いた表現が有効であることが確認できた。

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

加茂憲一, 福井敬祐, 坂本亘, 伊藤ゆり. がん対策立案・評価における意思決定に寄与するマイクロシミュレーションの構築：大腸がんを事例に, 計量生物学, 41, 93-115, 2021.

2. 学会発表

なし

H. 知的財産権の出願・登録状況

該当なし