次に，作成ボタンを押すことで，最適化に必要な情報を算出設定ファイルの情報を読み取り，新たに，最適化設定ファイルを作成・保存し，そして，図８に示すように，中央部のメモコンポーネントに，最適化設定ファイルが読み込まれる．この項目には，ユーザー自身が計算環境によって編集し直すべきものがいくつかある．同一人ペア率が小さいものには関心がない場合もあるかも知れない，その場合には，「ListUp Minimum PSame100」の値を大きくしても良い．デフォルトでは，０．００００１としている．この値によって，最終的に得られる同一人ペア率を記述したファイルのサイズが変わってくる．今回の設定では，２５万ペアのレコードがリストアップの対象となり，「RPL _ABS001 _S1 _PSAME.txt」として出力され，ファイルサイズは２３MB であった．例えば，この値を０とすれば，すべてのレコードペアが出力され，そのレコードペア数は１４億を超え，かなりのファイルサイズになる．

次に，「Parameters Prob for Same Person」の項目で，同一人ペアにおけるフィールドの一致率の最適化範囲を指定する．例えば，生年月日の月，YY では，同一人ペア率の初期値を０．９０，その最適化範囲を０．６００から０．９９９に設定している．多くのフィールドは，同一人であれば，ほぼ一致するであろう項目のはずなので，同一人におけるそのフィールドの一致率は，０．９００から０．９９９の間で最適化すれば十分と思われる．

最適化の範囲を事前情報に基づいて狭めることも可能である．例えば，【００】または【A０】で設定した既知結合キーが存在する場合，【A１】一致型ペア数タブには，一番右端に，結合済みの同一人ペア数が表示される．この値を用いて，同一人ペアにおけるフィールドごとの一致率が算出が可能である．結合済みペア数は合計１７８３ペアである．その中で，生年月日の年，YY のみ不一致となったペア数は８件ある．したがって，同一人ペアにおいて，YY が一致する確率は，（１－８÷１７８３）＝０．９９５と計算できる．実際，後ほど，【B１】最適化の結果タブに，YY の同一人ペアにおける一致率が０．６００から０．９９９の範囲で最適化されているが，その値は０．９９５７８となり，既知結合済みのレコードペアによる０．９９５とかなり近い．また，同様にして，MM，DD,UNI,NA の既知結合済みレコードペアによる一致率の概算値と最適化値（括弧内）は，それぞれ，０．９９６（０．９９６），０．９７６（０．９７５），０．９９０（０．９５７），０．９８７（０．９８８）となり，どれもかなり近いことが分かる．このように，事前に，もしくは，前回の照合結果がある場合には，これを用いて，最適化の範囲をせばめることができる．

最後に，追加データファイルにおける同一人ペアレコードの率（Proportion of Same Person）の設定を行う．これは，初期値のみの設定で，ここでは，０．１となっており，追加データファイルの１割の人が，基本データベースに同一人に関するレコードを持つことを想定している．この値は，ユーザーごとに大きく変える必要があるかもしれない．しかしながら，すでに，照合を行った経験があれば，どの程度の同一人に関するレコードがあるか容易に想定

できる．

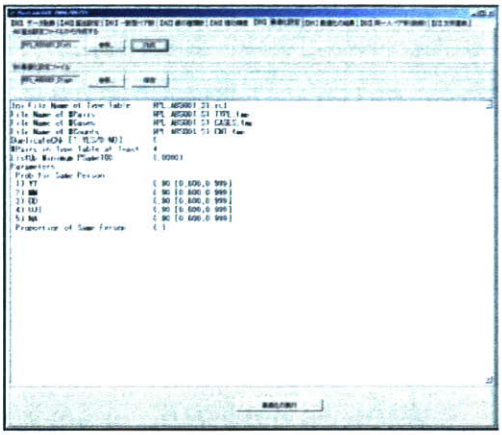


図８．最適化の設定

【B0】最適化設定タブにおいて，最適化設定ファイルの編集を終え，上部の「保存」ボタンを押せば，変更が保存される．そして，底部の「最適化の実行」ボタンを押すと，図９のようなコンソールアプリケーションが起動し，最適化の計算が始まる．この計算は非常に多くの時間を必要とし，何回かの繰り返し計算を行いながら，最適化する．今回のデータでは，CPU：INTEL T2500　@２．００GHｚ，RAM：２．０GBのノートPCを利用して，約８０分かかった．計算は何回かの繰り返しを行うが，１単位ごとに，その残り時間や，経過時間を表示するのみで，全体の計算時間などを予測しない．このコンソールアプリケーションは，計算終了と同時に，自動的に終了し，閉じる．

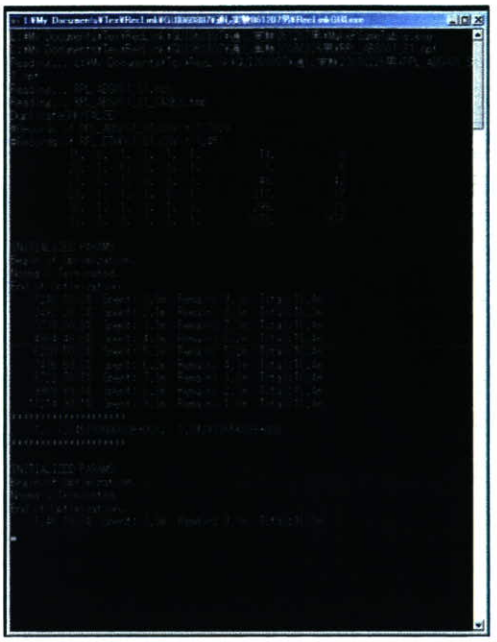


図９．最適化の計算

繰り返し計算の収束状況は，図１０に示すように，「RESULT.tmp」ファイルとして，保存されている．

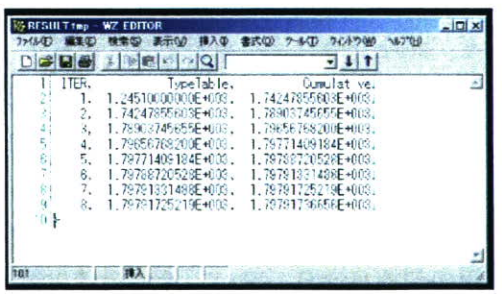


図１０．繰り返し計算の収束状況

最適化が終了すると，【B１】最適化の結果タブ，【B２】同一人ペア率（抜粋）タブに，その結果が表示される．まず，図１１に示される，【B１】最適化の結果タブの表示項目について説明する．同一人に関するレコードペアにおけるフィールドごとの一致率の推定値は，Prob for Same Person に

表示される．そして，別人ペアにおけるフィールドごとの値の種類数，すなわち，とり得る値の場合の数は，【A２】値の種類数タブでの結果が表示されている．次に，追加データファイルにおける同一人レコード率が，０．１４，すなわち１４％と表示されている．そして，すべてのレコードペアにおける同一人ペア数の推定値が，１７９７．９１人となっており，そのうち，【B０】最適化設定の「ListUp Minimum Psame100」で指定した値以上のペアに含まれる同一人ペア数が，１７９７．８８人と示されている．既存の結合済みペア数が今回の例では分かっており，１７８３人であったので，かなり良い精度で，同一人ペアを探していると言える．また，リストアップの対象からもれた同一人ペア数は，０．０３人と算出されている．【B０】最適化設定の「ListUp Minimum Psame100」を高く設定すれば，取りこぼされる同一人ペア数も多くなることに注意する必要がある．

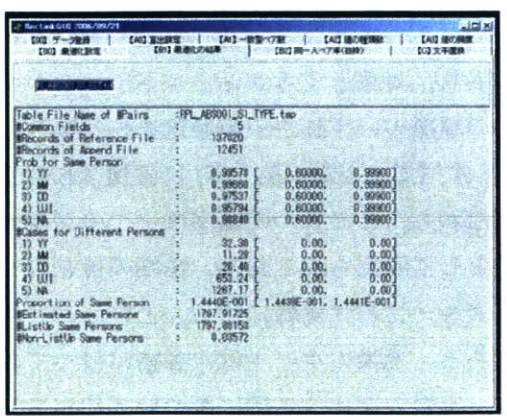


図１１．最適化の結果

【B２】同一人ペア率（抜粋）タブには，基本データファイルの個人識別番号，追加データファイルの個人識別番号，もしあれば，結合済みの個人識別番号，レコードペアごとに，各フィールドの一致・不一致の型，同一人ペアである確率，別人ペアである確率が付与されたファイルの抜粋が読み込まれている．最後の列は，フィールドの値の稀度合いが示されており，稀な漢字での名前の一致などの情報を集約したものである．この値によって，各フィールド値の一致・不一致の型が同じであっても，レコードペアごとに，同一人ペア確率が変わる可能性がある．ファイル名は，ここでは，「RPL _ABS001 _S1 _PSAME .txt」となっており，データファイルと同じフォルダに保存されている．

ユーザーは，この照合結果ファイル，RPL _ABS001 _S1 _PSAME .txt を駆使して，照合作業を行うことになる．テキストファイルのままでは，大きなデータの操作は難しいので，SQL 機能を持つデータベースソフトウェアのファイル形式に変換する必要がある．そして，同一人ペアの絞込みを試みる．手がかりになるのは，同一人ペア率（Psame100）が大きいもの，そして，【A１】一致型ペア数タブに表示された一致型ペア数の表になる．表には，一致型ごとの該当するペア数，また，結果ファイルには，同一人ペア率が記述されているので，データベースを利用して，一致型ごとに含まれる同一人ペア数を算出することも可能である．

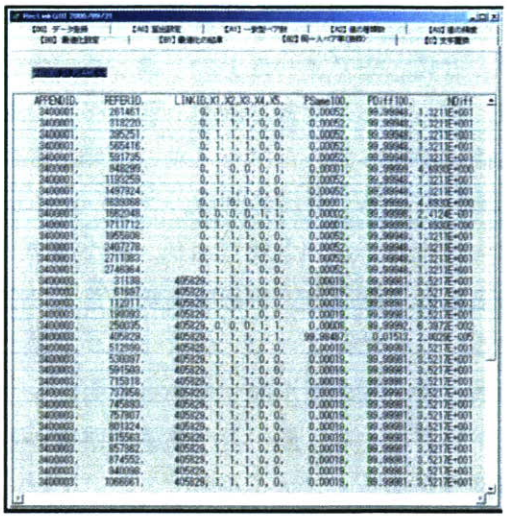


図１２．各レコードペアにおける同一人ペア率

この照合ソフトで生成されたファイルの一覧を図１３に示す．すべての解析は，照合の対象となる２つのデータファイルから始まっている．ソフトウェアの各タブに表示された表，計算結果はすべてテキストファイルとしてデータファイルと同じフォルダに保存されている．設定ファイルなどは，必要に応じて再利用可能であるし，どのような設定で解析を行ったかを後で確認すると言う意味でも保存しておく方が良い．

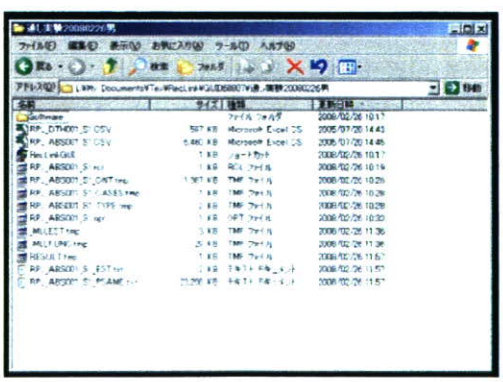


図１３．照合ソフトウェアによって利用または作成されたファイル

B４．ひらがば・カタカナ・漢字の変換

照合作業の前処理として，ひらがな，カタカナ，漢字の変換が考えられる．すなわち，記入されたデータファイルにも，タイピングミス，転記ミス，思い違い，などから，間違った文字で，氏名などが記述されることも少なくない．そこで，ある程度，間違いやすい文字，古い漢字，同音のカタカナなどを適当な文字に変換することが考えられる．ここでは，【C】文字変換タブの機能を簡単に説明する．

文字変換は，指定したデータファイルの指定した列，フィールドについてのみ行われ，その度に，名前を付けてファイルを保存する必要がある．置換の作業は設定ファイルとして保存され，再利用可能である．置換のルールはソフトウェアと同じフォルダに含まれるテキストファイル「全置換表.txt」に記述してあるため，ユーザーの扱うデータファイルの特性を考えて，適当に編集しなおす必要があるかも知れない．例えば，このファイルの中に，相似がある漢字のグループとして，「李 季 秀」という行があり，対象とするフィールドに，この３つの単語のいずれかの値が含まれる場合には，すべて，左端の値，「李」に置換される．すなわち，この３つの漢字は一つのグループとして考えられており，左端の値を以ってグループの代表値にする，というルールである．変換のルールは，多岐に渡っており，実際の記述ミス例に基づいて作成されているが，すべてのルールを用いると変換し過ぎる可能性もある．繰り返すが，使用する前に，変換ルールが記述された，全置換表.txt を一度検討し，不必要なルールを削除してから適用すべきである．

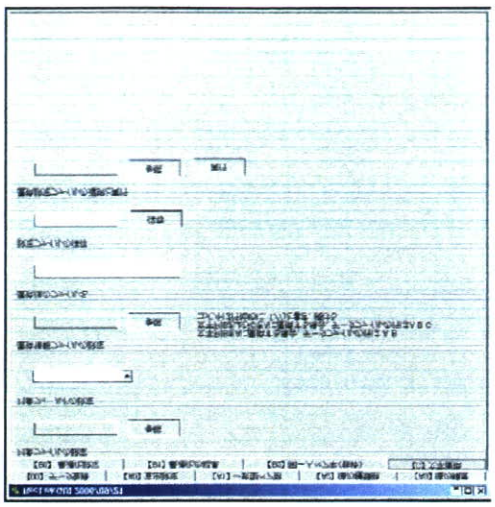


図１４．事前の文字変換

C. 研究発表

C1. 学会発表

1. Tonda T., Satoh K., Kawasaki H.*1, Shimamoto T., Nakayama, T*2, Katanoda K.*3, Sobue T.*3 and Ohtaki M. (*1Graduate School of Health Science, *2Oita Univ. Nurs. Health Sci., *3National Cancer Center): Statistical analyses of spatial-time distribution of cancer mortality in Japan, 18th annual meeting of the International Environmetrics Society, Mikulov, Czech Republic, 2007.
2. 冨田哲治, 鎌田七男*1, 大瀧 慈 (*1 広島大学名誉教授): 長期曝露の発がんリスクを評価するための統計モデル, 2007年度統計関連学会連合大会, 神戸, 2007.
3. 3. Tonda T., Satoh K., Kawasaki H.*1, Shimamoto T., Nakayama T.*2, Katanoda K.*3, Sobue T.*3, Sato, Y.*4, Yamaguchi, N.*4 and Ohtaki M. (*1Graduate School of Health Science, *2Oita Univ. Nurs. Health Sci., *3National Cancer Center, *4Toyko Women's Medical University): Statistical Analysis of Spatial-Time Heterogeneity of Cancer Mortality Risk Based on Growth Curve Model, East Asia Regional Biometric Conference 2007, Tokyo, 2007.
4. 冨田哲治, 佐藤健一, 川崎裕美*1, 島本武嗣, 中山晃志*2, 片野田耕太*3, 祖父江友孝*3, 佐藤康仁*4, 山口直人*4, 大瀧慈 (*1 医学部保健学科, *2 大分県立看護科学大学, *3 国立がんセンター, *4 東京女子医科大学): Statistical Analysis of Spatial-Time Heterogeneity in Cancer Mortality Risk, 科研費シンポジウム「バイオインフォマティックスおよび経時観察データの解析」, 広島, 2008.

C2. 論文発表

1. Izumi S.*1 and Ohtaki M. (*1Oita Univ.): Incorporation of inter-individual heterogeneity into the multistage carcinogenesis model: Approach to the analysis of cancer incidence data. *Biomedical Journal* **49**(4), 539-550, 2007.
2. 小田光子*1, 佐藤健一, 岸田典子*2 (*1 比治山大学, *2 安田女子大学): 小児生活習慣病予備軍の簡易スクリーニング手法の開発による要予防検診者や要生活指導者の判別について, 栄養学雑誌, 65-5, 233-240, 2007.
3. Hiyama E., Iehara T., Sugimoto T., Fukuzawa M.,

Hayashi Y., Sasaki F., Sugiyama M., Kondo S., Yoneda A., Yamaoka H., Tajiri T., Akazawa K., Ohtaki M.: Screening at 6 months of age reduced mortality of euroblastoma: A retrospective population-based cohort study including more than 13 million Japanese screened infants, Lancet (in press).

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

がん死亡動向分析および地理分布解析

分担研究者 水野正一 国立健康・栄養研究所 生物統計プロジェクト

研究要旨

いくつかのがんの部位が地域差を示し注目されている。結腸のがんは、北に高い特徴が見いだされ、これを、国民栄養調査の府県別エネルギー摂取量と関連させてみると、1980年代の報告では、エネルギー摂取量は北に高く、それは、1990年代の府県別 BMI と相関が高く注目された。BMI の値自体は平均値として通常域内での変動であるが、今後分散成分の検討、通年変動と関連させて検討を深める必要が示唆された。

A．研究目的

いくつかのがんの部位が地域差を示し注目されている。結腸のがんは、北に高い特徴があり、Vit-D と関連させ、紫外線照射量との関連が議論されている。今回、エネルギー摂取量、ＢＭＩとの関連を検討した。

B．研究方法

資料と方法：人口動態指定統計目的外使用の許可のもと年齢各歳ごとのがんの部位別年度別（1972～2002）死亡数を用いた。分母の人口は５年ごとの国勢調査時各歳人口を用いた。

エネルギー摂取量、ＢＭＩの地域差は、国民栄養調査報告をもとにした府県別の報告値を用いた。 1975～2000 年の結腸のがんの年代別年齢都道府県別の死亡率をもとに SMR を算出し地域相関を検討した。

C．研究結果

1. 結腸のがんの都道府県別 SMR を算出し図に示した（図 1）

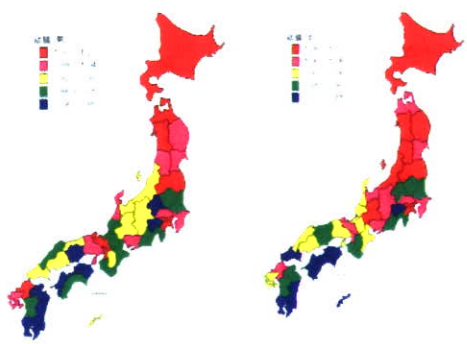


図 1. 結腸のがん(SMR)地理分布

これをみると東京都、大阪府、北九州、愛知県等の大都市を含んだ地域に SMR 値が高い一方で、北の方のいくつかの県で SMR が高い傾向が見いだされた。

府県別のエネルギー摂取量の地域分布に関しては、国民栄養調査を基にしての 1980-84 年 の Yamaguchi らの都道府県別の検討、1995-1999 年の検討が存在する。エネルギー摂取量の都道府県別結果を図に示した（図２）

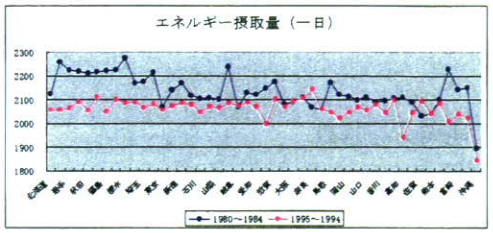


図２．都道府県別一日エネルギー摂取量

これをみると、 1980-84 年の結果 では、北に高いという特徴が見いだされた。1995-1999 年の報告では、エネルギー摂取量は低下し、全体的な一様さが特徴であった。 1995-1999 年次の BMI の地域分布報告を図３に示した。

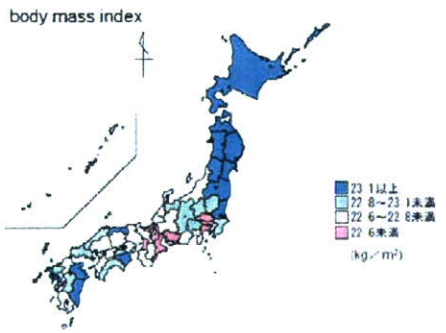


図 3. 1995-1999 年次の BMI の分布

図３においては、北の方のいくつかの道県において、BMI が 23.1 以上のカテゴリーとして報告されていて、結腸のがんの死亡率の「北に高い」という部分に関して、一致をみせた。

D．考察とE．結論

結腸がんは、近年増加傾向にあって、その動向が注目されている。食事摂取、肥満度（ＢＭＩ）との関連は大きいとされている。地域分布に関しては、US Data にて北に高いという報告があって、Vit-D との関連で紫外線日射量の関連が議論されている。（参考図４）

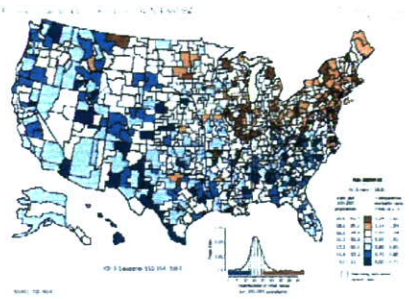


図４．Colorectal Cancer SMR White-male 1988-92

結腸のがんは、日本においても「北に高い」ということがあって、同じように紫外線日射量の関連は議論されている。今回、国民栄養調査資料をもとにして都道府県別の報告を頼りに、エネルギー摂取量、ＢＭＩの地域差の検討を加えた。Yamaguchi らの 1980-84 年の結果では、エネルギー摂取量は、「北に高い」という特徴があって注目された。1995-1999 年の検討ではエネルギー摂取量は全体として 1980-84 年当時よりは低めの値が報告（北のほうでは特にそうである）され、地域差も少なくなった印象を与えた。これには、国民栄養調査が、一日調査になったことの影響はどうかとか調査法の変遷も考慮する必要があるのかもしれない。

一方 BMI（1995-1999）の地域分布は、「北に高い」という点においては一致していて注目された。しかしながら、平均値のカテゴリーが、22.6 未満～23.1 以上というなかでの違いであり、ＢＭＩの地域差の結腸のがんの寄与に関しては今後の検討が必

要である。肥満の流行に関しては、４月から保健指導が新しく制度として実施されることもあって、地域差に関しての理解が進むことが期待されている。

F．健康危険情報

特になし

G．研究発表

１．論文発表

1. M Tomita, S Mizuno, K Yokota: Increased levels of serum uric acid among ex-smokers. （ accepted and will appear in JE 2008 ）

2. 水野正一, 渡邊昌：糖尿病トクホの問題点 (will appear in Fucntional Food 2008).

２．学会発表

1. 水野正一、片野田 耕太、祖父江友孝：がん死亡動向分析および地理分布解析. 第30回がん疫学研究会 2007/07/13 東京.

H．知的所有権の取得状況

１．特許取得 なし

２．実用新案特許　なし

３．その他　なし

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

がん罹患の動向に関する研究

分担研究者　加茂　憲一　札幌医科大学講師

研究要旨
現在報告されている全国がん罹患数は，地域がん登録における登録の完全性と，全国数推定方法に起因して過小評価されている可能性が指摘されている．この問題に対し昨年度は，非線形回帰モデルを用い，登録の完全性を補正した罹患数を推定する方法を提案した．今年度は，この方法の数理的な改良を行い，2001年の全国値を推定した結果を報告する．

A．研究目的

地域がん登録は，がんの実態を表す代表的な指標である罹患・死亡・生存率のうち，罹患と生存率を把握するために必須の仕組みである．また，様々な医療技術の発展によりがんの予後が改善されてきた今日，がんの１次予防効果を反映する「罹患」は重要な指標である．罹患の国レベルの動向を把握するために必要不可欠な情報が，全国がん罹患数推定値であるが，本研究では，全国がん罹患数の推定方法に着目した．

現在報告されている全国推定値は，過小評価の可能性が指摘されている．その原因の１つは，地域がん登録における登録の完全性の低さである．完全性の程度を計る指標として従来用いられてきたのは，DCN(Death certificate notification)割合・DCO(Death certificate only)割合・IM(Incidence, mortality)比の３つである．登録が完全な状態では，DCN割合やDCO割合は0%になるはずだが，20～30%程度の地域が非常に多いのが，我が国の現状である．この点に関しては，地域がん登録の早急な全国的整備が必要である．

全国がん罹患数推定値が過小評価されるもう１つの原因は，その推定方法に起因する．上記のように地域がん登録における登録の完全性の問題がある事が分かっているため，推定には以下の２条件を両方ともクリアした地域のデータのみを用いている：

条件１：IM比が1.5以上
条件２：DCN割合が30%未満，またはDCO割合が25%未満

この「足切り」により，ある程度完全性の高い地域のデータのみを用いて推定を行う事になるが，この足切りをクリアした地域でさえDCN割合が20%の地域が多いのが我が国の現状である．更に，このステップ以外では，登録の完全性に関する補正は全く行われない．従って，この手順で推定された全国数は，DCN割合が約20%での全国数とみなす事が出来る．地域がん登録においてDCN割合が低くなれば，自動的に罹患数は増加することを考えると，全国がん罹患数は地域がん登録の完全度に依存して変化することになり，これでは真の罹患を把握出来ていない．

そこで昨年度は，非線形回帰モデルを用いて，DCN割合が0%における罹患数を推定するモデルを構築した．このようなモデルによって登録が完全な状態の罹患数を推定する試みは初めてであったので，昨年度提案したモデルは多くの仮定を要する１stステップである．今年度は，昨年度のモデルに改良を加えた２ndステップと呼べるモデルを提案し，全国罹患数報告の最新年である2001年の推定を行い，報告数との比較を行った．

B．研究方法

昨年度は，地域がん登録における登録の完全性を補正した全国がん罹患数推定方法として，IM比とDCN割合の間にある非線形関係を用いた回帰モデルに基づく推定方法を提案した．具体的には，IM比をy，DCN割合をx，DCN割合が0%におけるIM比をβ（未知）とすると

$$y = \beta / (1 + (\beta - 1)x) \qquad \square \text{（式1）}$$

なる関係があることを用いて，パラメータβを推定する．このβの推定値に，全国がん死亡数を乗じる事により，DCN割合が0%における全国罹患数を推定するという方法を提案した．ここで，我が国のがん死亡数に関しては，人口動態より完全なデータが得られる事を利用している．

しかし，この方法を適用する（（式1）を導出する）ためには，幾つかの仮定をが必要である．その１つとして「がん死亡率は，既登録群と未登録群において等しい」という仮定があるが，これ自然ではないと考えられる．そこで，既登録群と未登録郡における，がん死亡率の差異を表す未知パラメータをαとして，αとβ両方を推定するようなモデルを構築した．具体的には未登録群におけるがん死亡率は，既登録群よりもα倍高いとする．

地域 i（i=1, 2, □, N）におけるがん罹患数をn_i，がん死亡数をm_i，DCN割合をx_iとする．また，MI比（IM比の逆数）をy_iとすると，$y_i=m_i/n_i$である．がん死亡は，がん罹患の中から確率y_iで発生すると考えられるので，m_iは二項分布に従い

$$m_i | x_i \sim \mathrm{Bin}(n_i, y_i) \qquad \square \text{（式2）}$$

と表現出来る．もしa=1ならば，y_iは（式1）で表されるが，αが1でなければ，y_iは

$$\beta y_i^2 + (\alpha x_i - \beta x_i - 1) y_i + (1-\alpha) x_i = 0$$

□ (式3)

の２つの解のうち大きい方で表される（(式3) は変数やパラメータの仮定の下で相異なる実数解を持つ）．パラメータ α，β には $\alpha>0$，$\beta>1$ という制約があるので $a=\log\alpha$，$b=1+\log\beta$ とおき未知パラメータベクトル θ を $\theta=(a,b)'$ とおくと，尤度関数は

$$n\Sigma w_i[y_i\log\{y_i(\theta)/(1-y_i(\theta))\} + \log(1-y_i(\theta))]$$

となる．ここで n は考察対象地域における全罹患数 Σn_i であり，$w_i=n_i\div n$ である．つまり，未知パラメータ推定における地域ウエイトは罹患数で入ることになる．罹患数は，地域規模にある程度比例し，また登録の完全性が高ければ増える数であるので，これら２つを同時にみたすウエイトであると考えられる．

次に，今回のモデルでは新たなパラメータ α を導入したが，果たしてこのパラメータが本当に必要かどうかを判定しなければならない．すなわち，次の２つのモデル：

モデル①　：$\alpha=1$
モデル②　：$\alpha\neq1$

のうち，どちらが適切かを判断せねばならない．モデル①は，本質的に従来の（式1）と同値である．モデル①と②では，どちらが適しているかを判定するのには，交差検証法(Cross-validation)を用いる．具体的には，以下の値を各モデルにおいて算出し，小さい値を取る方を採択する：

$$CV = -2\Sigma\{m_i\log p_{i[-i]} + (n_i-m_i)\log(1-p_{i[-i]})\}$$

□ (式4)

ここで，$p_{i[-i]}$ は y_i のジャックナイフ推定量を表す．ジャックナイフ推定量とは，i番目の観測値を抜いて推定した値である．この規準量はStone(1974, 1977)により提唱された情報量規準である．

C．研究結果

上記の方法を2001年における15の地域がん登録データ（山形県，宮城県，千葉県，神奈川県，新潟県，福井県，愛知県，滋賀県，大阪府，鳥取県，岡山県，佐賀県，長崎県，熊本県，沖縄県における，罹患数，死亡数，DCN数）に適用し，2001年の全国推定を行い，現在報告されている数との比較を行った結果を表１に示す．

表１　2001年全国がん罹患数推定結果

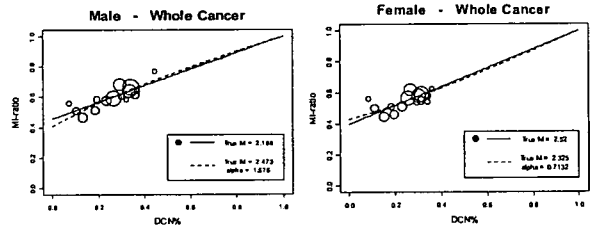
部位	性別	選ばれたモデ	報告罹患数	新たに推定された罹患数
全部位	男	①	325213	396761
	女	①	243568	300550
胃	男	①	72267	88087
	女	①	35459	42953
肺	男	①	49427	56209
	女	②	21192	34837
結腸	男	②	36582	78591
	女	①	29213	37408
直腸	男	②	21557	38787
	女	②	12785	22135
肝臓	男	②	27727	27037
	女	①	12745	14337
乳房	女	②	40675	60178
前立腺	男	①	23548	39983

新たに推定された全国がん罹患数は約70万人であり，報告されている数よりも約13万人多かった．別の表現をすると，実際には現在報告されている数の約1.2倍の罹患数が存在する可能性がある．

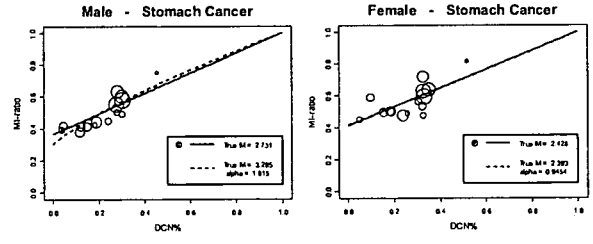
では，新たに推定された値(回帰曲線)と，実際のデータとの関係を散布図で見てみよう．全部位と胃がん，肺がん，直腸がんに関する実測値プロットと推定された回帰曲線を図1に示す．ここで，胃がんは最も罹患数が多く，かつ男女共にモデル①が選択された部位，肺がんは最も死亡数が多く，男女で選ばれたモデルが異なる部位，直腸がんは男女共にモデル②が選択された部位である．

図１　IM比－DCN割合プロットと回帰曲線

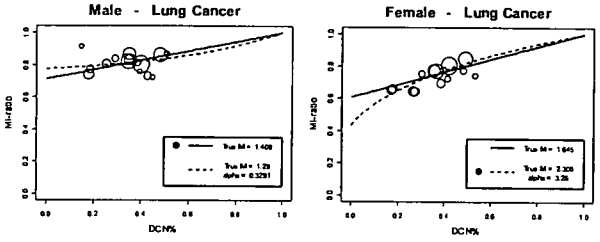
全がん



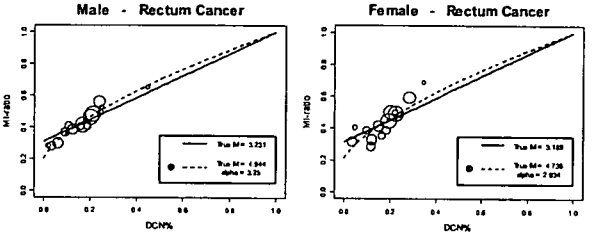
胃がん



肺がん



結腸がん



ここで，横軸はDCN割合，縦軸はMI比，プロットの大きさは地域の罹患数に比例する．また実線と破線はそれぞれモデル①，②における回帰曲線（モデル①においては回帰直線）を表す．また右下には，各モデルでの推定値を示し，CV規準量により選択されたモデルは◎印で示す．

D．考察

全国がん罹患数推定の新たな方法を提案した．その方法は，地域がん登録における登録の完全性を補正した推定値を得るためのものである．具体的には，昨年度報告の非線形回帰モデルを用いた解析方法に対する改良を行った．この改良によって「がん死亡率が，既登録群と未登録群で等しい」という仮定を外すことが可能となった．また，従来は推定における地域ウエイトを「人口」で入れていたが，これが「罹患」となり，地域規模と登録の完全性の２つを兼ね備えたものとなった．

今回の方法には２つのモデルを設定し，どちらが適しているかを判断するというステップが存在する．このうちの１つは，従来型であり，シンプルなモデルである．結果的には多くの部位で従来型のモデルで良い事が分かった．しかし，複雑なモデルが選択された部位も存在した．具体的には女性の肺がん，男性の結腸がん，男女共に直腸がん，男性の肝臓がん，乳がんである．このモデルにおいては，がん死亡率の差異を表すパラメータが導入されている．その推定値は男性の肝臓がん以外の全てで１より大きかった．この結果は，既登録群と未登録群の間にがん死亡率の差異があるときには，未登録群におけるがん死亡率の方が高い事を意味する．一方，男性の肝臓がんにおけるパラメータαの推定値は0.03と，非現実的な値であった．殆どの地域でDCN割合が40％と非常に高く，同じDCN割合でもIM比に大きなバラツキが存在するので，このような異常な推定値が得られたと考えられる．

昨年度，本年度と２年に渡って全国がん罹患数推定方法，特に登録の完全性を補正する方法に関する研究を行ってきた．真の罹患数の動向を知ることは，効果的ながん対策を画策する上で必要な事である．そのために，数理的なアプローチを行ってきた．しかし，このような方法によって，真の罹患数が完全に把握出来たと考えるのは早計である．望ましいのは，完全なデータに基づく推定であり，そのためには地域がん登録の全国的な整備が急務であろう．

G．研究発表

1．論文発表

1) Kamo K, Kaneko S, Satoh K, Yanagihara H, Mizuno S, Sobue T. A mathematical estimation of true cancer incidence using data from population-based cancer registries. Jpn. J. Clin. Oncol. 37: 150-155, 2007.

2) Marugame T, Katanoda K, Matsuda T, Hirabayashi Y, Kamo K, Ajiki W, Sobue T. The Japan cancer surveillance report: incidence of childhood, bone, penis and testis cancers. Jpn. J. Clin. Oncol. 37: 319-323, 2007.

3) Marugame T, Matsuda T, Kamo K, Katanoda K, Ajiki W, Sobue T. Cancer incidence and incidence rates in Japan in 2001: based on data from 10 Population-based cancer registries. Jpn. J. Clin. Oncol. 37: 884-891, 2007.

2．学会発表

1) 加茂憲一，柳原宏和，片野田耕太，松田智大，丸亀知美，味木和喜子，祖父江友孝：非線形回帰モデルによる全国がん罹患数推定，2007年度統計関連学会連合大会，神戸，2007.

2) 加茂憲一，丸亀知美，片野田耕太，松田智大，平林由香，味木和喜子，祖父江友孝：Comparison of method estimating nation-wide cancer incidence. 第66回日本癌学会学術総会，横浜，2007.

3) 片野田耕太，丸亀知美，松田智大，平林由香，加茂憲一，味木和喜子，祖父江友孝：Incidence pattern of soft tissue sarcoma in Japan - form population-based cancer registry in 1993-2002. 第66回日本癌学会学術総会，横浜，2007.

4) 加茂憲一，片野田耕太，松田智大，丸亀知美，味木和喜子，祖父江友孝：短期予測によるタイムリーな全国がん罹患数報告．第18回に本疫学会学術総会，東京，2008.

H．知的財産権の出願・登録状　（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

厚生労働科学研究補助金（第3次対がん総合戦略研究事業）
分担研究報告書

がん死亡率のAge-Period-Cohort分析

研究協力者　雑賀公美子[1)]、中村　隆[2)]、片野田耕太[1)]
分担研究者　味木和喜子[1)]
主任研究者　祖父江友孝[1)]
1）国立がんセンターがん対策情報センター　がん情報・統計部
2）統計数理研究所

研究要旨

1958-2006年のがんの死亡率の動向を、中村のベイズ型 Age-Period-Cohort モデルを用いて年齢、時代、世代の3効果に分離し、年次推移や出生年別の表示ではとらえることのできないがん死亡率の変動を検討した。年齢階級別の死亡率は、1980年代後半以降は高齢ほど高いが、1980年代後半までは75-79歳、80-84歳にピークがあり、85歳以上では死亡率が低い。年齢階級別死亡率を時系列でみると、75歳以上の高齢群では1958-2006年にかけて増加しており、75歳未満では逆に減少している。これらの死亡率の変動を3効果に分離すると、高齢ほど高い死亡率は年齢効果として取り出され、近年の85歳以上の急激な死亡率増加は、1890-1910年生まれの高い世代効果として取り出された。また、近年の男性75歳未満、女性80歳未満の死亡率減少も、交互作用としてではなく1900年以降世代効果が継続的に減少している傾向として取り出された。年齢×時代の交互作用効果は、70歳以上で1960年代は平均的な年齢効果を上回っていたのに対し1980年代以降は下回っており、これは70歳以上での死亡率が高い傾向は緩和してきていることを示している。時代効果に示された1970年代半ば以降の上昇は、がんの死亡率の上昇に応じて脳血管疾患の死亡率が減少しているなどの競合リスクの影響で、1995年の時代効果の上昇は、国際疾病分類（ICD）のICD-9からICD-10への改正における影響である。

今回はがんの全部位をまとめて分析したが、がんは部位によって危険因子が明らかにことなるため、世代効果の影響の説明などはやはり部位別の解析が必要である。また、死亡は競合リスクや死亡診断の影響などによる時代の効果が強調されるため、罹患分析との比較検討も必要である。

A. 研究目的

現在わが国の死因第1位であるがん死亡率のこれまでの動向を把握し、将来の動向を推計することは、がん対策の評価および対策の立案において重要な意味をもつ。本研究では、人口動態統計によるがん死亡データ（1958-2006年）を、ベイズ型 Age-Period-Cohort モデルを用いて年齢、時代、

世代の 3 効果に分離し、年次推移や出生年別の表示ではとらえることのできないがん死亡率の変動を検討することを目的とする。

B. 研究方法

1. 資料・対象

全がん死亡数は、人口動態統計[1]による 1958-2006 年の年別、年齢 5 歳階級別を用いた。人口は、5 年ごとに実施されている国勢調査人口および国勢調査時の人口を基準に算出されている推計人口を用いた。40 歳以上を対象とし、性別に分析を行った。

2. 年齢階級別死亡率動向の検討

がんの死亡率の動向を観察するため、年齢階級・調査年別の死亡率を 3 次元グラフに示した。便宜的に表示の年齢は各年齢階級の中央の年齢とした。

3. Age-Period-Cohort 分析

中村のベイズ型ポワソン Age-Period-Cohort モデル（BAPC モデル）[2][3]によって、3 効果を推定した。j 年の第 i 年齢階級の人口を P_{ij}、がん死亡数を d_{ij} とすると、BAPC モデルは以下のようなモデルである。

$$d_{ij} : \text{Poisson}(\lambda_{ij}),$$

$$\log \lambda_{ij} = \log P_{ij} + \beta^G + \beta_i^A + \beta_j^P + \sum_{k=1}^{K} w_{ij,k}\beta_k^C$$

ここで、β^G は総平均効果であり、β_i^A、β_j^P、β_k^C はそれぞれ年齢、時代、世代効果のパラメータである。$w_{ij,k}$ は j 年の第 i 年齢階級に対応する世代に対して世代効果パラメータ β_k^C がどの程度寄与しているかを表すウェイトであり

$$w_{ij,k} \geq 0, \quad \sum_{k=1}^{K} w_{ij,k} = 1$$

を満たす。各効果パラメータは、足して 0 になるようなゼロ和制約

$$\sum_{i=1}^{I} \beta_i^A = \sum_{j=1}^{J} \beta_j^P = \sum_{k=1}^{K} w_{ij,k}\beta_k^C = 0$$

を課して基準化しておく。

APC モデルには、「(世代) = (時代) − (年齢)」という関係があるため、何らかの付加条件がなければ、年齢、時代、世代の 3 効果の値が一義に決められないという識別問題が存在する。中村はこの識別問題を克服するために、各効果のパラメータの 1 次階差の 2 乗和を最小にする漸進的変化の条件を与えるモデルを発案した。

また、本研究では年齢と時代の交互作用効果 β_{ij}^{AP} をもつ以下の BAPC モデル[3]も検討した。

$$\log \lambda_{ij} = \log P_{ij} + \beta^G + \beta_i^A + \beta_j^P + \sum_{k=1}^{K} w_{ij,k}\beta_k^C + \sum_{i_*=1}^{I}\sum_{j_*=1}^{J} w_{ij,i_*j_*}\beta_{i_*j_*}^{AP}$$

$$w_{ij,i_*j_*} \geq 0, \quad \sum_{i_*=1}^{I}\sum_{j_*=1}^{J} w_{ij,i_*j_*} = 1$$

$$\sum_{i_*=1}^{I}\sum_{j_*=1}^{J} w_{ij,i_*j_*}\beta_{i_*j_*}^{AP} = 0$$

2 つのモデルを比較し、赤池のベイズ型情報量規準 ABIC（Akaike's Bayesian Information Criterion）が最小になるモデルを選択した。

C. 研究結果

1. 年齢階級別死亡率動向の検討

図 1、図 2 に男女のがん死亡率の動向を示した。男性ではどの調査年でも 40 歳以降で高齢になるほど死亡率は高い。1958-72 年までは 75-79 歳の死亡率がもっとも高く、80 歳以上は低い。1973-84 年では 80-84 歳

の死亡率がもっとも高く、1985 年以降は85 歳以上の死亡率がもっとも高い。年齢階級別の時系列変化では 1995 年までは 75 歳以上の年齢階級で年々高くなっており、85 歳以上では上昇が顕著、75 歳未満ではあまり変化はない。1995 年以降では 75 歳以上ではあまり変化なく、75 歳未満では 2006 年にかけて死亡率は低くなっている。

女性では 1958-68 年では 75-79 歳の死亡率が、1969-80 年では 80-84 歳の死亡率が、1981 年以降は 85 歳以上の死亡率がもっとも高い。時系列変化は、1958-2006 年の全期間を通して年々上昇するのは 85 歳以上のみであり、75 歳未満は年々低下傾向、75-84 歳においても 1995 年以降は低くなっている。

2. Age-Period-Cohort 分析

男女とも、年齢と時代の交互作用を含むモデルは、交互作用効果をもたないモデルより ABIC の値が減少していたため（男性で 899.3、女性で 975.9 ほど減）、以降は交互作用効果をもつモデルについての結果を述べる。

図 3、図 4 に、それぞれ男性と女性の BAPC モデルの結果を示した。各図の 4 つのパネルは、左から順に時代効果、年齢効果と年齢×時代の交互作用効果、年齢効果と年齢×時代の交互作用効果を足したもの、世代効果の推定値をプロットしたものである。2 番目のパネルの太線は全期間を通しての平均的な年齢効果であり、細線は年ごとの交互作用効果を示している（上にいくほど最近年）。左から 3 番目の交互作用効果のパネルは、ある時代における年齢効果として解釈できるよう、年齢効果に交互作用効果を足したものである。

1） 年齢効果

男女とも年齢が高いほど年齢効果が高い。男女を比べると男性の方が変動は大きい。

2） 時代効果

男女とも 1970 年代半ばまでは大きな変動はないが、それ以降に上昇傾向が見られる。また、1995 年において急な上昇が観察された。

3） 年齢×時代の交互作用効果

男女とも、1960 年代は 70 歳以上の部分で全期間を通じた平均的な年齢効果を上回っており、1980 年代、1990 年代は逆に下回っている。さらに 2000 年以降では 60 歳代、70 歳代の効果が平均効果より低い。左から 3 番目のパネルが示す時代ごとの年齢効果は、男性より女性の方が直線的に増加する。

4） 世代効果

男女とも、1870 年代生まれから 1900 年生まれまでは上昇傾向、以降若い世代ほど世代効果は低下する。

D. 考察

がんの粗死亡率は、男性で 1978 年、女性で 1984 年以降死因の第 1 位であり、1980 年から 2006 年にかけて、全死亡に対するがんが占める割合は男性で 23.9%から 34.1%、女性で 20.6%から 26.1%と年々増加している [1]。この原因として、医療の進歩による脳血管疾患や心疾患による死亡の減少およびがんの有病者の増加、平均余命の長期化

によるがん罹患人口の増加などが考えられるが、様々な要因が混在する死亡率の年次推移をもとに要因を考察することには限界がある。全がん死亡率に年齢、時代、世代が与える効果を APC 分析により明らかにし、それぞれの効果の大きさの変化と、これらに対応する要因について検討した。

1) 年齢効果

年齢効果は、時代や世代に関係なく、生理的側面やライフステージと関連した加齢による変化を捉えたものである。

年齢効果が高齢ほど上昇することは、がん対策や予防に関係なく、高齢ほどがんの死亡率が高いことを示している。

2) 時代効果

時代効果は、年齢や世代に関係なく、社会全体が変化する部分を表したものであり、それぞれの時代背景や政策などによる社会環境の変化による影響を捉える。

1970 年代後半以降、時代効果が上昇傾向にあるのは、死因が「がん」の者が年齢に関係なく増加していることを示している。

本研究と同様の BAPC モデルを用いた脳血管疾患死亡率の分析によると、1970 年頃から時代効果は減少しており [4]、全死亡に占めるがん死亡の割合も男性で 1978 年、女性で 1984 年に脳血管疾患死亡を抜いて 1 位になっているなど、脳血管疾患死亡の減少ががん死亡の増加と関連していることが示唆される。死因診断による影響や競合リスクの影響を受けないがん罹患の APC 分析（1975-1994 年）では、今回の死亡の結果ほど大きな時代効果は見られない上、近年の増加傾向も観察されていない [5]。そのため、死亡の APC 分析結果で示された時代効果には、競合リスクの増減が大きく関連しており、罹患には見られない死亡特有の時代効果の上昇があることが明らかとなった。

また、0 歳時の平均余命の年次推移をみると、1958 年では男性 65.0 歳、女性 69.6 歳であったのが 2006 年には男性 79.0 歳、女性 85.8 歳と約 15 年増加している [6]。この平均余命の増加からも、がんに罹患したまま生存する人が増加しており、がんは治癒という概念が明確ではないため、がん罹患経験のある患者の死因が「がん」とされる傾向があることも否定できない。

1995 年に観察された急な時代効果の上昇は、国際疾病分類（ICD）の ICD-9 から ICD-10 への改正において、ICD 死因分類の変更、死因選択ルールの変更、死亡診断書の変更により死因ががんとされる死亡例が増加したためと考えられる。

3) 年齢×時代の交互作用効果

2000 年以降の 60 歳代、70 歳代の年齢と時代の交互作用効果が平均的な年齢効果を下回っていることに加え、1960 年代の 60 歳未満の交互作用効果も平均効果を下回っていることから、この年齢の死亡率が高い傾向が緩和してきていることが示されている。

4) 世代効果

世代効果は、年齢や時代による変化以外の、生まれ育った生活環境やそれまで生きてきた時代背景により得た、他の世代と区別できる特徴を表したものである。

図 1、図 2 の 3 次元グラフにみられた年

齢階級別の死亡率が 1970 年代くらいまでは 80 歳以上または 85 歳以上の高齢群で低くなっているのは、1880 年代後半から 1910 年代生まれまでの世代効果が高いためと考えられる。また、近年の 75 歳未満に顕著な死亡率の減少傾向も近年の 1900 年以降生まれの世代効果が低下しているためと考えられる。

今回の分析結果で観察された 1980 年代後半から 1900 年生まれまでの世代効果の上昇傾向は、がん罹患の APC 分析でも観察されている。しかし、今回の死亡分析では 1900 年生まれ以降は継続した減少傾向が観察されているが、罹患分析では 1900 年生まれ以降 1930 年代後半生まれまでは減少傾向、1950 年代前半生まれで高い世代効果を示すものの、その後あまり変化はみられていない。このように世代効果が死亡と罹患で異なることの解釈は、生涯リスクの違いなどと関連してくると考えられるが、がんの危険因子は部位によって大きく異なるため、部位別の解析がさらに必要である。

E. 結論

がんの死亡率の近年の増加は、時代効果の上昇傾向と 1870 年代生まれから 1900 年生まれ（近年の高齢者）の世代効果が高いことで主に説明されることが明らかとなった。時代効果の上昇傾向をストップさせる、または減少させることががん対策の目標となるが、死亡診断書への死因記載や死亡コードの変更が時代効果に大きな変化を与えるため、がん罹患分析との比較を行いながらの考察が必要である。

F. 研究発表

1. 論文発表

なし

2. 学会発表

雑賀公美子, 片野田耕太, 祖父江友孝. 国民健康・栄養調査による喫煙者割合の Age-Period-Cohort 分析結果と将来推計. 第 18 回日本疫学会学術総会, 東京, 2008 年 1 月.

G. 知的財産権の出願・登録状況

特になし

参考文献

1) 厚生労働省大臣官房統計情報部. 人口動態統計, 平成 18 年上巻, 2008.
2) 中村隆. ベイズ型コウホート・モデル-標準コウホート表への適用-. 統計数理研究所彙報, 29: 77-97, 1982.
3) 中村隆. コウホート分析における交互作用効果モデル再考. 統計数理, 53: 103-132, 2005.
4) 三輪のり子, 中村隆, 成瀬優知, 大江洋介, 大野ゆう子. わが国における 20 世紀の脳血管疾患死亡率の変動要因と今後の動向. 日本公衆衛生学会誌, 53(7): 493-502, 2006.
5) 大野ゆう子, 村田加奈子, 中村隆, 津熊秀明, 味木和喜子, 大島明. ベイズ型ポワソン・コウホートモデルによる日本のがん罹患の将来予測. 厚生労働省がん研究助成金 地域がん登録精度向上と活用に関する研究, 平成 15 年度報告書: 87-96, 2004.
6) 厚生労働大臣官房統計情報部. 簡易生命表 平成 18 年. 厚生統計協会. 2007 年 11 月.

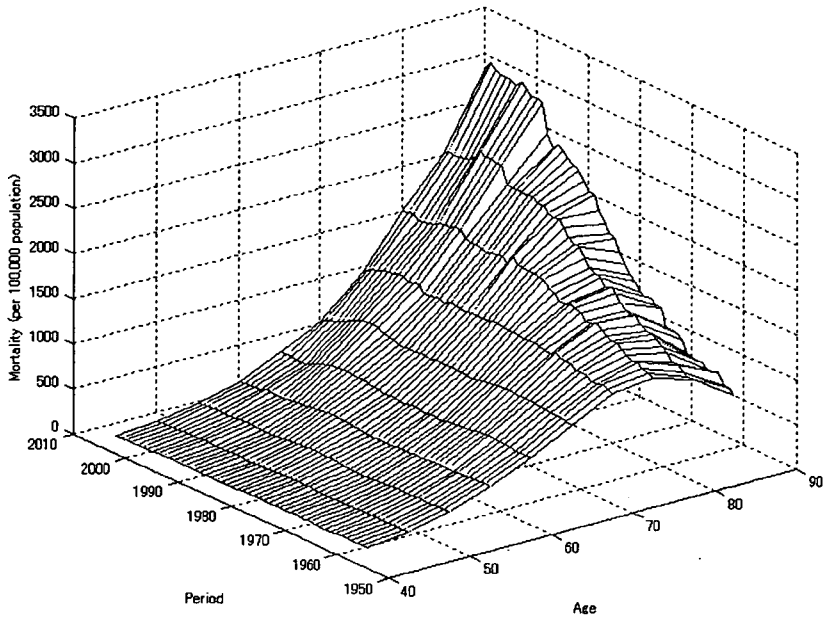


図 1. 全がん死亡率（男性）

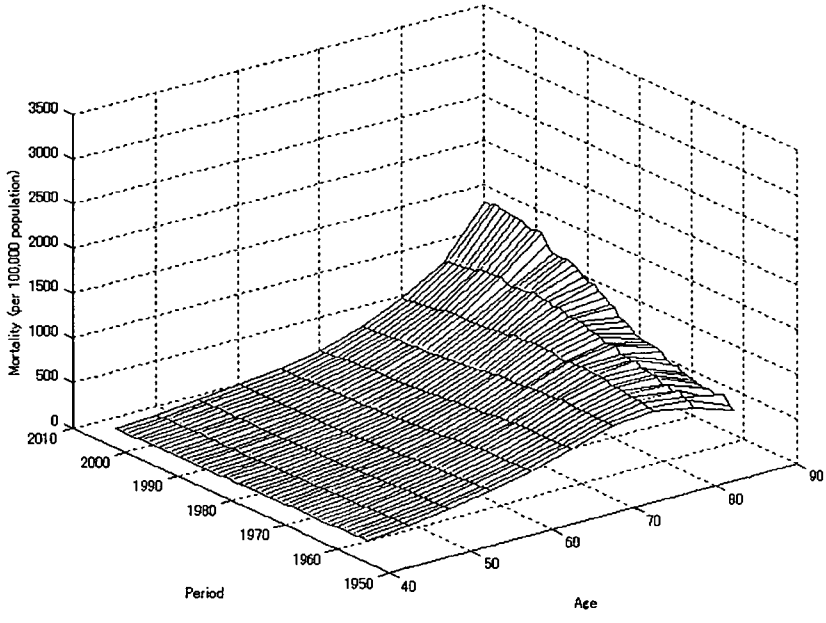


図 2. 全がん死亡率（女性）

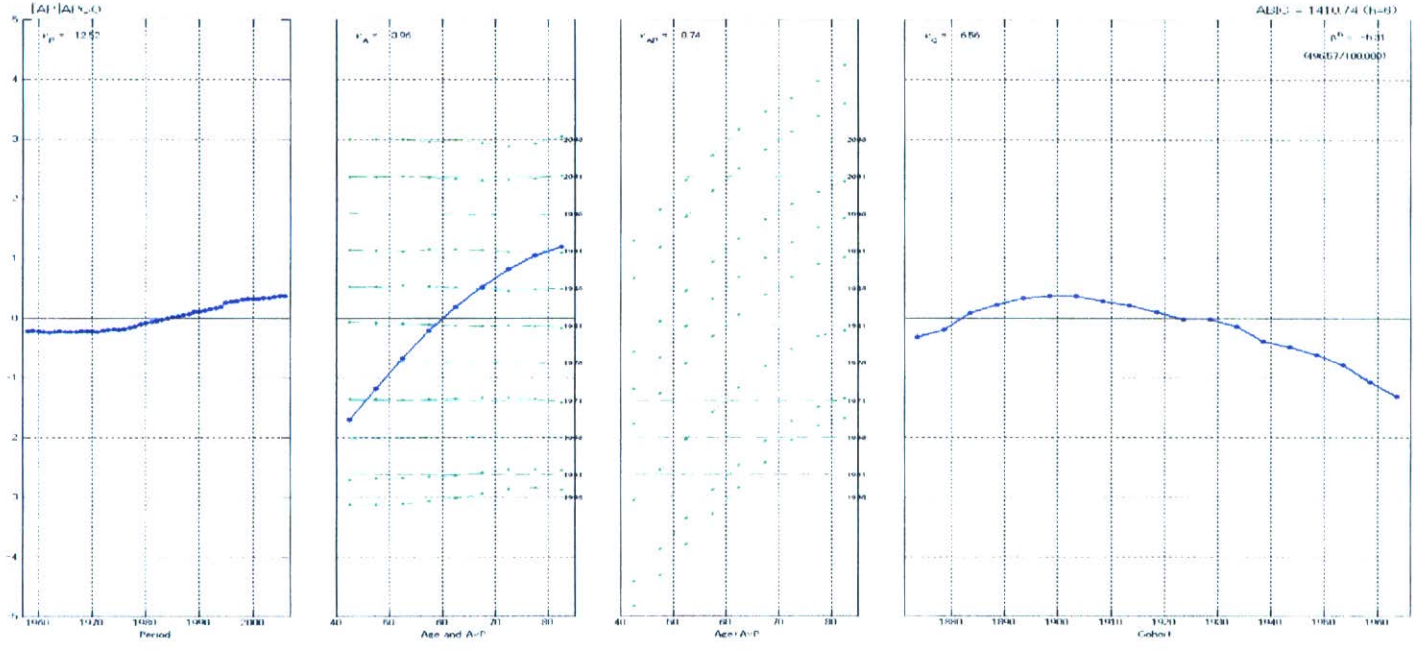


図 3. 全がん死亡率の APC 分析結果（男性）

1 枚目：時代効果

2 枚目：年齢効果と年齢×時代の交互作用効果（太線：全期間を通しての平均的な年齢効果、細線：年ごとの交互作用効果）

3 枚目：年齢効果と年齢×時代の交互作用効果を足したもの

4 枚目：世代効果

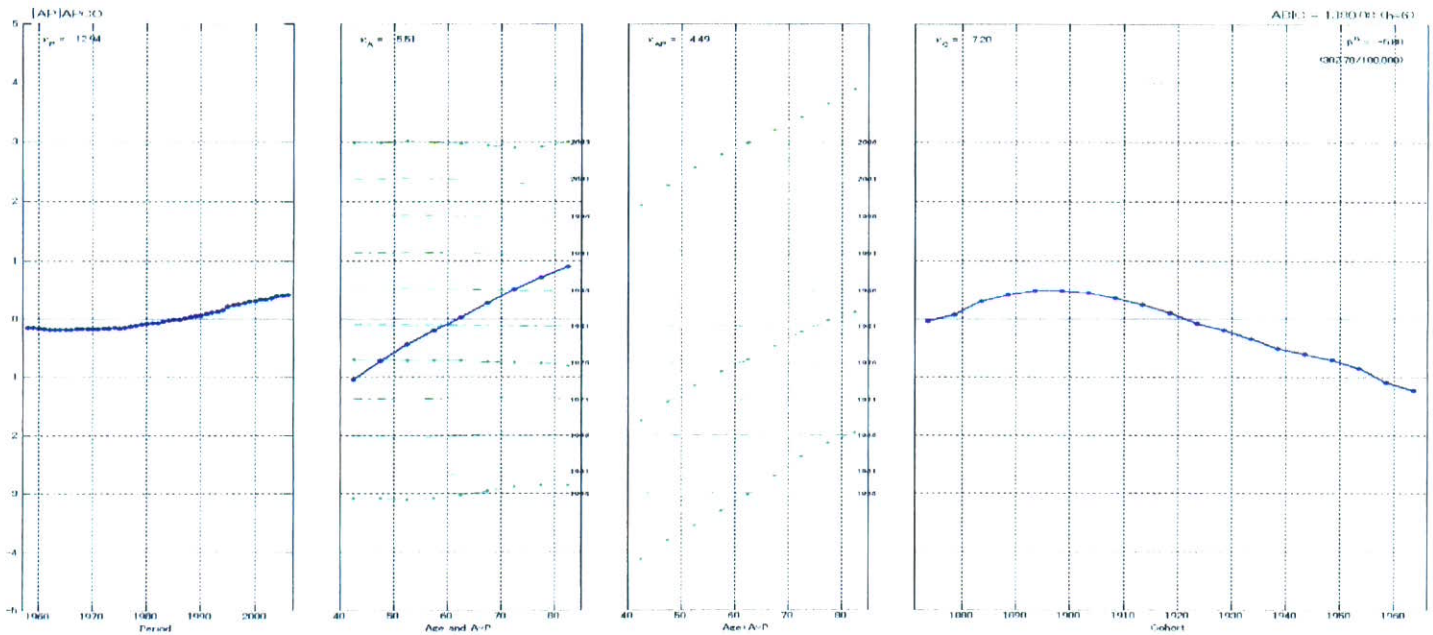


図 4. 全がん死亡率の APC 分析結果（女性）

1 枚目：時代効果

2 枚目：年齢効果と年齢×時代の交互作用効果（太線：全期間を通しての平均的な年齢効果、細線：年ごとの交互作用効果）

3 枚目：年齢効果と年齢×時代の交互作用効果を足したもの

4 枚目：世代効果