を行なった。

クライアント側で使用するデータ入力プログラムは、登録票、死亡票の入力、個人同定作業、これらの作業に伴う各種リストの作成機能を提供する。それぞれの票における入力時にはディクショナリー参照機能を使用でき、項目チェックも行なえる。また、複数の入力担当者による同時入力、同時同定作業も可能である。

D. 考察

地域がん登録標準システム開発における今後の課題として以下のものが挙げられる。(1)個人を同定するために、標準データベースシステムでは姓、名、生年月を同定指標として用いるが、大規模人口県においては、この指標のみでは候補が多数現れることから、より絞り込んだ同定作業を行なう必要がある。そのため、同定指標にがん登録部位、性など、他のどのような同定指標を組み合わせることが効果的であるのかを検証する。(2)死亡票から多重がんを登録するか否かについて統一的な見解を得た上で、それに基づいた入力の検討を行なう。死亡票エントリーテーブル（死亡小票に記載された内容がエントリーされる）において、１票の死亡票から複数腫瘍を入力可能にし、死亡票マスターテーブルに必要な項目を登録する。但し、罹患数に関わる死亡票からの集計ルールは標準的なものとして決定しておく必要がある。またこの点が決定後罹患数と登録精度にも影響する。これについて、課題として引き続き検討・決定すべきことは大阪成人病センター味木和喜子参事による資料を参照する。(3)「がん」として抽出する範囲の決定（死亡票記載欄、「腫瘍」の扱い、病脳期間）を行なわなければならない。(3)DCO定義の整理を行なわなければならない。これは、来年度着手予定の集計用（腫瘍単位に要約・集計）システム開発に関わる。

E. 結論

地域がん登録の登録精度と登録の即時性の維持・向上のために、本研究班で標準データベースシステムの構築が決定されたのを受け、放射線影響研究所情報技術部にてシステムの作成を行なった。この報告書では、標準データベースへの移行に関しての諸工程について記述した。

F. 健康危険情報

特になし

G. 研究発表

特になし

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

レコードリンケージにおける個人同定処理の自動化に関する研究

分担研究者　大瀧　慈　広島大学原爆放射線医科学研究所計量生物研究分野教授

研究要旨

レコードリンケージにおける個人同定処理は、計算機が未発達であった時期から必要とされていたこともあり、国内では経験的な手法についての議論が先行している。本研究では、経験的な知識の統計モデルへの利用方法も議論し、計算機による同定処理の自動化を行う際に有用なレコード値の頻度を考慮した判定基準を提案した。解析例によって、統計モデルの適合度を検証することができ、新規でレコードリンケージを行う場合のみならず、既存の照合結果の検証にも有効であることが示唆された。

A．研究目的

現在、多くのがんセンターでは、登録患者の生死および死亡要因の情報を、都道府県が有する死亡人口動態データを参照することで更新している。ところが、わが国には、個人識別番号がないために、登録患者と死亡者のレコードが同一人に関するものかどうかは、これらに共通の項目（氏名、年齢、生年月日、住所など）の一致状況によって判断するしかなく、経験的な手法を用いた人手による個人同定処理が行われている。

小規模のデータベースであれば、このような照合作業は人手によって対応可能であるが、がん登録など数万人規模のレコードに対しては、網羅的な照合は事実上不可能である。また、経験的手法による弊害として、どの程度の同定ミスがあるのか把握できず、その程度も施設ごとに異なる、という問題が挙げられる。例えば、登録者で実際には死亡したにも関わらず、同定から漏れてしまった場合には、異常な年齢になるまで、しばらくは、生存者として扱われる。このような登録者が多ければ、がん登録者全体の寿命が伸びることになり、データの信頼性が失われ、その有効利用が危ぶまれる。また、同様にして、照合手法の違いが、施設ごとの寿命の違いを生み出してしまう。これらの問題は数十年経たないと、表面化しないため、初期対応が非常に重要である。

それゆえ、すべての施設に共通な、個人同定処理の標準化は急務であり、また、この処理による誤判別の把握も重要な課題である。本研究では、より定量的な議論ができるように、統計モデルを用いたレコードリンケージを考え、個人同定処理の自動化に必要な理論とソフトウェアの整備を行う。

B．研究方法

理論の開発とソフトウェアの開発については以下の通りである。

① 理論の開発

異なるデータファイルからそれぞれ１つずつレコードを取り出すと、以下の表１のような一致型を得る。

表１．一致型の例.

項目名	氏	名	元号	年	月	日
$x^{(A)}$	佐藤	春夫	明治	4	4	9
$x^{(B)}$	佐藤	健一	昭和	4	4	9
$t(x^{(A)};x^{(B)})$	1	0	1	1	1	1

我々の目的は、与えられたレコードのペアが「同一人に関するレコードである」ことの確からしさを定量的に評価し、これに基づいてレコードリンケージを行うことにある。従来法では、一致型 t を持つペア集合から無作為に１つ取り出したレコードのペアが同一人ペアである確率(以後、同一人ペア確率)は、次式で評価されおり、この確率が高い一致型を持つレコードペアが同一人ペアの候補となっている。

$$r^{(S)}(t) = \frac{N^{(S)}p^{(S)}(t)}{N^{(S)}p^{(S)}(t) + (N^{(A)}N^{(B)} - N^{(S)})p^{(D)}(t)};$$

これまでの議論は、すべてのレコードペアを考えたものであったが、新しく得られたファイルの中の固定されたレコードと同一人ペアになるものを既存のファイルの中から探すという状況がより現実的である。そこで、ファイルＢのレコードを一つ固定し、ファイルＡの各レコードとの同一人ペア確率を考え、以下のレコード値の稀具合を反映させた同一人のペア確率を新たに提案した。

表 2. 固定されたレコード値とファイル A における頻度の例. ファイル A における頻度は男性のみを対象とし、$N_A = 138,594$.

項目名	氏	名	元号	年	月	日
$x^{(B)}$	佐藤	春夫	明治	4	4	9
$N_A \cdot f_k^{(D)}(x^{(B)})$	307.99	969.19	2.84	18.77	13.21	42.9
$x^{(B)}$	谷崎	潤一郎	明治	19	7	24
$N_A \cdot f_k^{(D)}(x^{(B)})$	2857.38	139694.00	2.84	40.58	13.91	41.43

すなわち、レコード$x^{(B)}$がデータファイルＡの別人に関するレコードと一致型 t を持つ確率を$p^{(D)}(t|x^{(B)})$とすれば、ファイル Ｂのレコード$x^{(B)}$を一つ固定したときの 同一人ペア確率は

$$r^{(S)}(t|x^{(B)}) = \frac{\alpha p^{(S)}(t)}{\alpha p^{(S)}(t) + (N^{(A)} - \alpha)p^{(D)}(t|x^{(B)})};$$

で評価可能である。実際、表２を見ても分かるように、別人において名が「春夫」で一致することは珍しくないが、「潤一郎」で一致することはかなり稀である。このように、珍しい名前で一致している場合、別人ペアで一致する確率は非常に小さくなり、相対的に、提案した同一人ペア確率は高くなる。

実際の照合処理においては、同一人ペア確率を昇順もしくは降順に並び替えることにより、自動判定処理、もしくは、人手処理の対象とすべき複数のレコードペアを選出するのが一般的である。 図１には、提案した同一人ペア確率が大きい方から同一人ペアの候補を選んだ場合の累積同一人ペア数を、図２には、その場合に誤って混入する別人ペア数を、そして、図３には、同一人ペア確率が小さいペアを別人ペアとした場合に誤って混入する同一人ペア数を示した。このように、新たに提案された基

準によって、自動処理を行った場合のご判別の把握が可能となる。

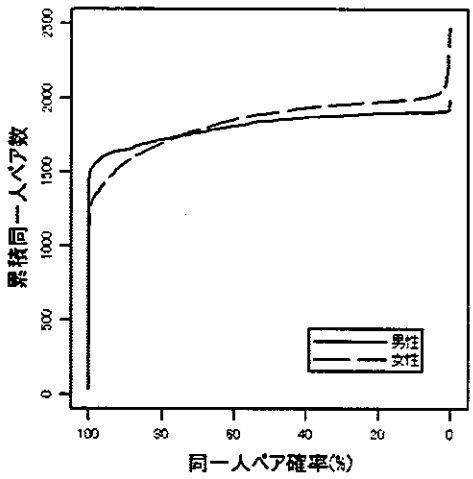


図 1. 同一人ペア確率と累積同一人ペア数
ファイル B のすべてのレコードに対して、ファイル A のすべてのレコードとの同一人ペア確率を算出した。横軸に同一ペア確率 $10^{\hat{a}}$ を、縦軸に区間 $I^{\hat{a}} = [\hat{a}; 1]$ に含まれる同一人ペア数 $q^{(S)}(I^{\hat{a}})$ を与えた。

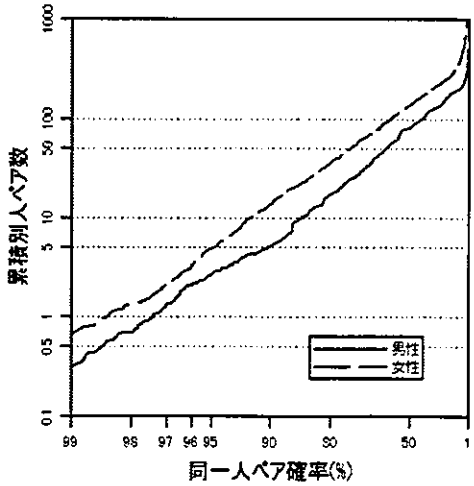


図 2. 自動判定における誤判別制御のための累積別人ペア数. ファイル B のすべてのレコードに対して、ファイル A のすべてのレコードとの同一人ペア確率を算出した。横軸に同一ペア確率 $10^{\hat{a}}$ を、縦軸に区間 $I^{\hat{a}} = [\hat{a}; 1]$ に含まれる別人ペア数 $q^{(D)}(I^{\hat{a}})$ を与えた。

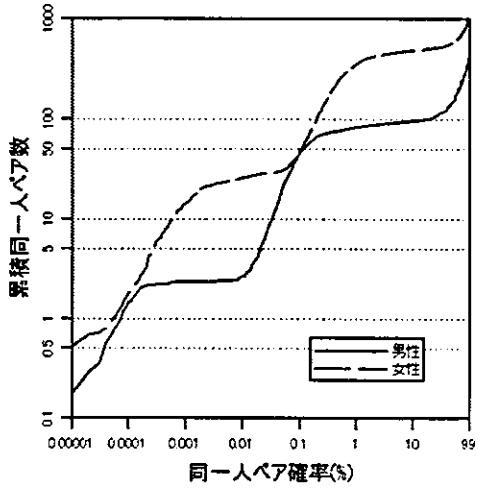


図 3. 自動判定における誤判別制御のための累積同一ペア数. ファイル B のすべてのレコードに対して、ファイル A のすべてのレコードとの同一人ペア確率を算出した。横軸に同一ペア確率 $10^{\hat{a}}$ を、縦軸に区間 $I_{\hat{a}} = [0; \hat{a}]$ に含まれる同一人ペア数 $q^{(S)}(I_{\hat{a}})$ を与えた。

（倫理面への配慮）

ソフトウェアの開発および実施テストには、原爆放射線医科学研究所が所有する暗号化されたデータを使用し、個人識別情報を直接使用することは避けた。

C．研究結果

本研究で開発したソフトウェアを用いて既に照合処理を行っている原爆放射線医科学研究所が所有するデータ（30 万レコー

ド）と県関連のデータ（3 万レコード）の個人同定処理を男女別に行った。その結果、ソフトウェアによって同一人ペア割合が高いと推定されたペアについては、ほとんど人手によって同一人と判断されたものであった。逆に、ソフトウェアによって同一人ペア割合が低く推定されたものであっても、人手によって同一人であると判断されたものがあった。また、少数例ながら、ソフトウェアによって新たに同一人に関するレコードとして同定されたペアもあった。

D. 考察

ソフトウェアによって同一人ペア割合が低いと推定されたものであっても人手によって同一人ペアと判断されたケースは興味深く、データの管理者と検討した結果、二つの原因が判明した。一つ目は、実際の照合作業では一人１レコードではなく、病歴などを含め複数個のレコードがあり、同一人として決め手となる情報が他にも存在した。さらに、共通項目以外の情報の中にも有用なものがあった。二つ目は、転記ミスなどの容認である。例えば、「アキコ」と「あきこ」や新旧漢字の違い、住所などの略記がこれにあたる。ソフトウェアでは、文字列として完全に一致しているか、そうでないかしか見ていなかったために、このような常識的な範囲内の曖昧な判断ができなかった。しかしながら、過去の個人同定例を参考にすることで、許容され得る値についてデータベースを作成することは可能であり、ソフトウェア的に改良の余地はあると思われる。

E. 結論

本研究では，個人同定処理のための理論の開発とこれをコンピュータ上で実現するためのソフトウェアの開発を行った。実際のデータを用いて実施テストを行った結果、本ソフトウェアで、同一人ペア割合が高いと判定された場合には高い信頼性を持って同一人ペアであると判断できることが示された。また、新たに提案した稀具合を反映した同一人ペア確率によって、個人同定の自動処理を行った場合の、精度が制御できるようになった。なお、男性よりも女性の方が一致型に仮定すべき確率構造が複雑であり、独立な確率では対応が困難であることが示唆された。新たに判明した問題点については、理論的およびソフトウェア的観点から、今後もさらなる検討が必要である。

F. 健康危険情報　特になし

G. 研究発表

１. 論文発表　なし

２. 学会発表

1) 佐藤健一・早川式彦・隅田治行・大瀧　慈. A statistical method for automatic identification in record linkage, International Biometric Conference, 2004 年 6 月, ケアンズ, オーストラリア.

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

アニメーション地図による最近の日本における がん死亡危険度の時空間分布の視覚化

分担研究者　大瀧　慈　広島大学原爆放射線医科学研究所・計量生物研究分野・教授

研究要旨

厚労省から目的外使用を許可された人口動態統計票のうち1975年から2002年までの28年間における全国の市区町村別死亡数，および1975年から2000年までの期間に行われた6回の国勢調査による市区町村別性別年齢階級別人口数を用いて主要部位のがんについて，市区町村別死亡数データに対してポアソン・ガンマモデルに基づく経験ベイズ法およびノンパラメトリック平滑化を適用し，各年次毎の性別市区町村別死亡相対危険度の推定値を算出した。得られた死亡相対危険度の推定値は，人口規模の小さい町村の場合にはその近傍の時空間平滑値に近似するが，人口の大きい都市では，近傍での状況に関わらずその都市独自の死亡相対危険度を反映するという特性を持っている。主要ながんの部位別に推定された市区町村別死亡危険度を５個のカテゴリーに層別化して地図上にプロットすることで，時空間分布の視覚化を行い、アニメーション化した。その結果、結腸がん，乳がん，脳腫瘍の死亡危険度は，この期間全国的に急増していること，肺がんは都市部を中心に全国的に徐々に増加していること，肝臓がんは西日本から東日本に向かって高死亡危険度地域が拡大していたが，1990年頃以降頭打ち状況にあること，その一方で，胃がんと子宮がんは全国的に急減していることが分かった。また，それぞれの部位のがんの死亡危険度が、時空的に局所的な危険度分布を持っていることが明らになった。

A．研究目的

この研究は，がんの部位別死亡危険度の時空間分布を年次別市区町村別人口動態データに基づいて推定するための統計理論の開発およびコンピュータアルゴリズムの作成することを目的としている。 いま，を第　市区町村の第　年次での観察死亡数とし，その帰無仮説の下での期待値をとするとき，その標準化死亡比は，

$$SMR_{it} = \frac{y_{it}}{\cdot_{it}}$$

となる。この標準化死亡比は，いわゆる直接法による指標に較べて人口の影響を受けにくいという利点を持っているが，人口数が小さな町村の場合には，観察死亡数が０や１を中心にした離散値であること，および　　が０に近い値を持つことにより，分散が大きくなり挙動が不安定となり，この指標による死亡危険度の時空間分布の地図

イメージ化は困難となる。

この問題に対して，これまで幾つかの対処法が提案されている。ベイズモデルの適用により SMR の推定値の安定化を行う方法として，Clayton and Kaldor (1987), Tsutakawa (1988), Tango (1988), Kenneth et al. (1989) は，ポアソン・ガンマモデルに基づく経験ベイズ法の適用について検討している。Xia et al. (1997) は，MCMC 法を用いた階層的ベイズモデルの適用による解析法を提案している。Kibria et al. (2002) は多次元ガウス分布を応用した階層的ベイズモデルに基づく空間分布の把握を試みている。また，SMR の空間平滑化を行うことで，尤もらしい地理分布を抽出する方法が提案されている。Liaw and Hwang (1997) は，台湾における市区町村別肺がん死亡の SMR データに対して，半径として全ての市区町村に関して共通な値を最適化した円近傍を各市区町村毎に定め，局所平滑化処理を適用して，死亡危険度地図を作成している。今回我々は，経験ベイズ法とノンパラメトリック平滑化法を組み合わせた新しい手法を考案した。なお，我々の空間平滑化では，円近傍として全国共通な半径を持つものの代わりに，それぞれの近傍内の人口数が共通となるように各市区町村ごとに可変な半径を持つものを使用し，各がんの部位毎にその人口数の値の最適化を行った。

B．研究方法

データは，厚労省から目的外使用を許可された人口動態統計票のうち 1975 年から 2000 年までの 25 年間における全国の市区町村別死亡数，および 1975 年から 2000 年までの期間に行われた 6 回の国勢調査による市区町村別性別年齢階級別人口数である。

[モデル]

従来，観察死亡数の分布に対しては，ポアソン分布を想定することが一般的であるが，今回は，潜在的な超過分散を考慮して，経験ベイズモデルの一種であるポアソン-ガンマモデル（Tsutakawa et al. (1985); Tango(1988)）の適用を行った。即ち，第市区町村における第

年次での死亡数は，下記の平均を持つポアソン分布に従うことを想定した，

$$\mu_{it} = \xi_{it}\mu_t\rho_i(t),$$

$i = 1,\ldots,n, t = 1,\ldots,k$.

ここで，μ_t は年次tにおける全ての市区町村の平均的相対死亡危険度，$\rho_i(t)$ は年次tでの第i市区町村の全国平均に対する相対死亡危険度を表す母数であり，死亡危険度の地理分布のコントラストの指標となるものであり，

$$\rho_i(t) = \lambda_{it}z_{it},$$

と表されるものとする。ただし，λ_{it} は(i,t)の時空間近傍における平均的な相対死亡危険度の対数値を表す固定効果母数であり，z_{it}は，(i,t)における相対死亡危険度の変動効果を表す母数で，その事前分布は，平均1，分散σのガンマ分布に従っているものとする。また，市区町村毎に時間的近傍で集計した死亡数 $\sum_{r\in N_i(\delta)} y_{ir}$ および期待死亡数$\sum_{r\in N_i(\delta)} \xi_{ir}$ も使用する。以下，y_{it} やξ_{it} と区別して使用するために，それぞれを $y_{i[t;\delta]}$ および$\xi_{i[t;\delta]}$ と記す。また，年次数

と市区町村数については,以下のモデルや推定量の表記における一般性を持たせるために,k および n と記すことにする。

[未知母数の推定]

Step1　全国単位の年次別平均的 SMR の算出

最初に，下記の式を用いて全国単位の年次別の平均的 SMR を算出する，

$$\widehat{\mu_t}=\frac{\sum_{i=1}^{n} y_{it}}{\sum_{i=1}^{n} \xi_{it}},\ t=1,\ldots,k.$$

全国単位の場合には，人口が多いため観察値と期待値の双方ともある程度以上の数となりそれらの比として定められるの値はそのままでも十分に滑らかな年次変動を持つようになるはずである。もしも，滑らかさが十分でない場合には，ノンパラメトリック平滑化処理を適用して，尤もらしい年次変動を抽出し，その結果を，と記して 全国単位の平均的 SMR とする。

Step 2　時空間近傍における平均的な相対死亡危険度の対数値を表す固定効果母数に対する推定

いま，2つの市区町村 i と j の役場の緯度と経度が与えられるとき，これらの役所間の距離 x_{ij} を，地球を球とみなして近似的に求める。この距離に基づいて，第 i 市区町村での半径 r の円近傍における相対死亡危険度の平均を算出する際の第 j 市区町村に関する重みを

$$w_{ij}(r)=\begin{cases} c_i\left\{1+\cos\left(\dfrac{\pi}{r}x_{ij}\right)\right\}, & 0\le x_{ij}\le r, \\ 0, & x_{ij}>r, \end{cases}$$

により定める。ただし，c_i は重みの総和が 1 となるための係数である。

いま，$r_i(M)$ を第 i 市区町村の円近傍のうち人口数が M を超える最小のものの半径の値とする。

Step F1.　次式で定義される CV 型規準量 $Q(\delta,M)$ を最小とする (δ,M) を求める，

$$Q(\delta,M)=\frac{4}{nk}\sum_{t=1}^{k}\sum_{i=1}^{n}\left\{\sqrt{y_{it}}-\sqrt{\xi_{it}\frac{\sum_{j\neq i} w_{ij}(r_i(M))y_{j[t;\delta]}}{\sum_{j\neq i} w_{ij}(r_i(M))\xi_{j[t;\delta]}}}\right\}^2-1,$$

Step F2. 前 Step で最適化された (δ,M) を $(\hat{\delta},\hat{M})$ と記すとき，各年次別に各市区町村近傍での相対死亡危険度の重み付き空間平均を次式により算出する，

$$\widehat{\lambda_{it}}=\frac{\sum_{j=1}^{n} w_{ij}(r_i(\hat{M}))y_{j[t;\hat{\delta}]}}{\sum_{j=1}^{n} w_{ij}(r_i(\hat{M}))\xi_{j[t;\hat{\delta}]}}\widetilde{\mu_t}^{-1},$$

$i=1,\ldots,n, t=1,\ldots,k.$

Step F3 市区町村毎の $\widetilde{\lambda_{it}}$ $(t=1,\ldots,k)$ について，ノンパラメトリック平滑化を行う。

Step R1.　年次t毎に，変動効果の分散母数 σ_t に関する尤度関数

$$L(\sigma_t|(\bar{\lambda}_t))=\prod_{i=1}^{n}\frac{\Gamma(y_{i[t;\delta]}+\sigma_t^{-1})}{y_{i[t;\delta]}!\Gamma(\sigma_t^{-1})}\left(\frac{\widetilde{\mu}\xi_{i[t;\delta]}\bar{\lambda}_t}{\widetilde{\mu}\xi_{i[t;\delta]}\bar{\lambda}_t+\sigma_t^{-1}}\right)^{y_{i[t;\delta]}}\left(\frac{\sigma_t^{-1}}{\widetilde{\mu}\xi_{i[t;\delta]}\bar{\lambda}_t+\sigma_t^{-1}}\right)^{\sigma_t^{-1}},$$

を定め，その値を最大化する最尤推定値 $\widehat{\sigma_t}$ を求める。

Step R2. $\{\widehat{\sigma_t} \mid t=1,\cdots,k\}$ に対するノンパラメトリック平滑化を行い，その結果を，$\{\widetilde{\sigma_t} \mid t=1,\cdots,k\}$ と記す。

Step R3. 年次別に市区町村毎に，変動効果母数 z_{it} の事後平均値を次式により推定する，

$$\widehat{z_{it}} = \frac{y_{i[t;\delta]} + \widetilde{\sigma_t}^{-1}}{\xi_{i[t;\delta]}\widetilde{\mu_t}\widetilde{\lambda_{it}} + \widetilde{\sigma_t}^{-1}} = \frac{SMR_{i[t;\delta]} + (\widetilde{\sigma_t}\xi_{i[t;\delta]})^{-1}}{\widetilde{\mu_t}\widetilde{\lambda_{it}} + (\widetilde{\sigma_t}\xi_{i[t;\delta]})^{-1}},$$

ただし $SMR_{i[t;\delta]} = y_{i[t;\delta]}/\xi_{i[t;\delta]}$ である。

Step 4. 修正 SMR 値の算出

既述の手順で求められた固定効果母数（時空間平滑化対数相対死亡危険度）と変動効果母数の推定値を用いることで，次式により市区町村別年次別相対死亡危険の推定値が導かれる，

$$\widehat{\rho_i(t)} = \widetilde{\lambda_{it}}\widetilde{z_{it}} = \widetilde{\lambda_{it}}\frac{SMR_{i[t;\delta]} + (\widetilde{\sigma_t}\xi_{i[t;\delta]})^{-1}}{\widetilde{\mu_t}\widetilde{\lambda_{it}} + (\widetilde{\sigma_t}\xi_{i[t;\delta]})^{-1}}.$$

$$\widehat{SMR_{it}} = \widetilde{\mu_t}\widehat{\rho_i(t)} = \widetilde{\mu_t}\widetilde{\lambda_{it}}\frac{SMR_{i[t;\delta]} + (\widetilde{\sigma_t}\xi_{i[t;\delta]})^{-1}}{\widetilde{\mu_t}\widetilde{\lambda_{it}} + (\widetilde{\sigma_t}\xi_{i[t;\delta]})^{-1}}.$$

人口数の大きい都市の場合，が満たされるので，$\widehat{SMR_{it}}$ の値は通常の SMR に値に近くなる。一方，人口の少ない町村の場合には，$(\widetilde{\sigma_t}\xi_{i[t;\delta]})^{-1}$ の値が大きくなるために，$\widehat{SMR_{it}}$ の値は その町村の近傍に関する時空間平滑値 $\widetilde{\mu_t}\widetilde{\lambda_{it}}$ に近くなる。

5．時空間分布の視覚化

主要ながんの部位別に推定された市区町村別相対死亡危険を 5 個のカテゴリーに層別化して，1.20 以上の場合「赤色」，1.05～1.20 の場合「黄色」，0.95～1.05 の場合「緑色」，0.80～0.95 の場合「水色」，0.80 以下の場合「青色」を対応させ，地図上にプロットし地理分布の視覚化を行った。人口規模の大きさを地図表示に反映させるために，正方形のマークのサイズを 1985 年時点の人口数にその面積が比例するように設定した。こうして作成された地図イメージを gif 形式の画像ファイルとし，年次の順番で連続表示することにより，死亡危険度の時空間分布をアニメーションとして視覚化した。

（倫理面への配慮）

本解析においては、個人識別情報を使用していない。

C．研究結果

がん死亡危険度の時空間分布を特定するために、部位別に、粗 SMR、空間平滑化 SMR、ランダム効果、地域間コントラスト SMR および時空間最適化修正 SMR を推定し、それらに基づく各種アニメーション地図を作製した。図 1～図 5 に肺がん死亡に関する結果を例示する。

その結果、結腸がん，乳がん，脳腫瘍の死亡危険度は，この期間全国的に急増していること，肺がんは都市部を中心に全国的に徐々に増加しているが一部地域で減少し始めていること，肝臓がんは西日本から東日本に向かって高死亡危険度地域が拡大していたが，1990 年頃以降頭打ち状況にある

こと，その一方で，胃がんと子宮がんは全国的に急減していることが分かった。また，それぞれの部位のがんの死亡危険度が，時空的に局所的な危険度分布を持っていることが明らかになった。

D．考察

日本におけるがん死亡危険度は、その部位によって時空間分布が大きく異なっていることについて、提案する手法を使用して初めてその詳細が明らかにすることができたものと思われる。がん死亡の予防策を講ずる際に、部位別の死亡危険度における全国的な経年変動および任意年次での相対的地域差を観ることで、より高い効果を持つ対処法の探察へ繋がるものと思われる。

E．結論

がん死亡危険度の時空間分布を高い精度で効率よく推定するための新しい統計的手法を提案した。さらに最新のコンピュータグラフィックス技法を適用して、その詳細をアニメーション地図として視覚化するシステムを開発した。

F．健康危険情報

特になし

G．研究発表

1．論文発表

1) Hirokazu Yanagihara, Megu Ohtaki. A family of regression models having partially additive and multiplicative covariate structure, Bulletin of Informatics and Cybernetics, in press.

2) Kenichi Satoh, Hirokazu Yanagihara and Ohtaki, Megu. Clustering method by connected neighborhoods and its application. Advances and Applications in Statistics, 4(2), 223-231, 2004.

2．学会発表

1) Ohtaki, Megu. Statistical Method for Estimating Spatial-Time Distribution of Mortality based on Municipality-Specific Demogra p hic Data　第 8 回中国日本統計学シンポジウム、桂林（2004 年 10 月）

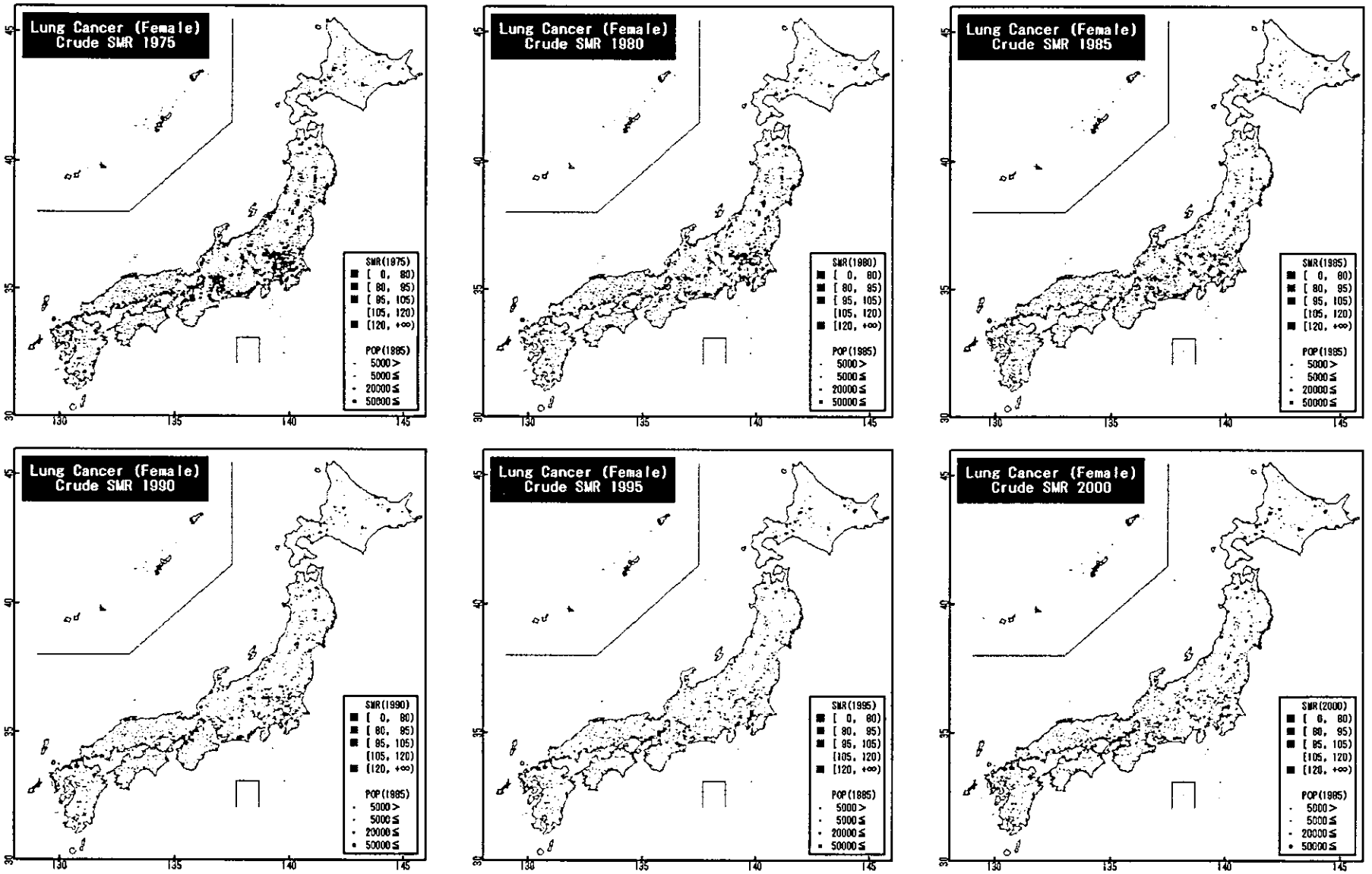


図1．女性の肺がん死亡に関する粗 SMR（ただし、1975 年～2002 年の全国を基準集団とする）アニメーション地図の一部抜粋

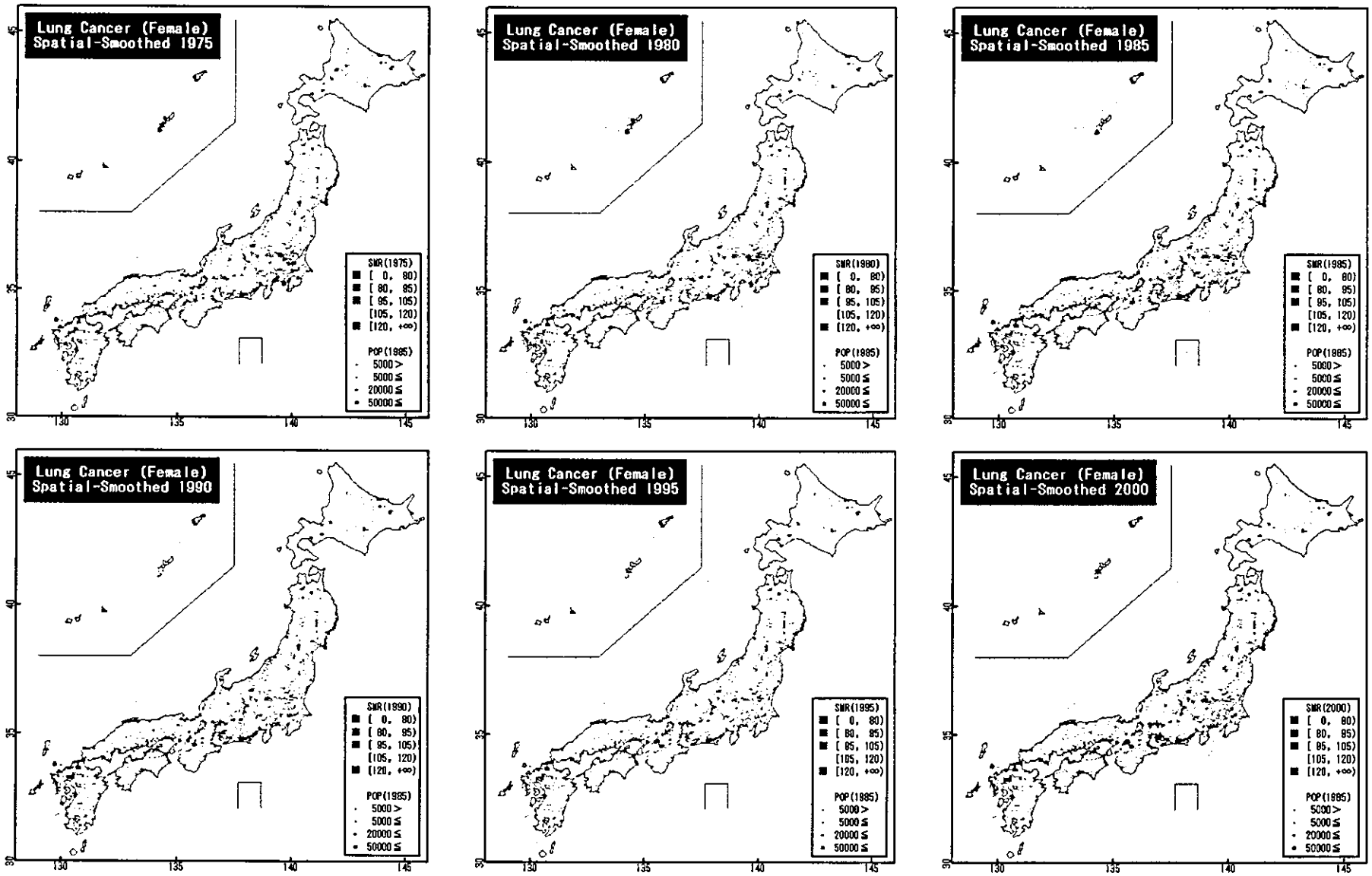


図2．女性の肺がん死亡に関する空間平滑化SMRアニメーション地図(1975年～2002年)の一部抜粋

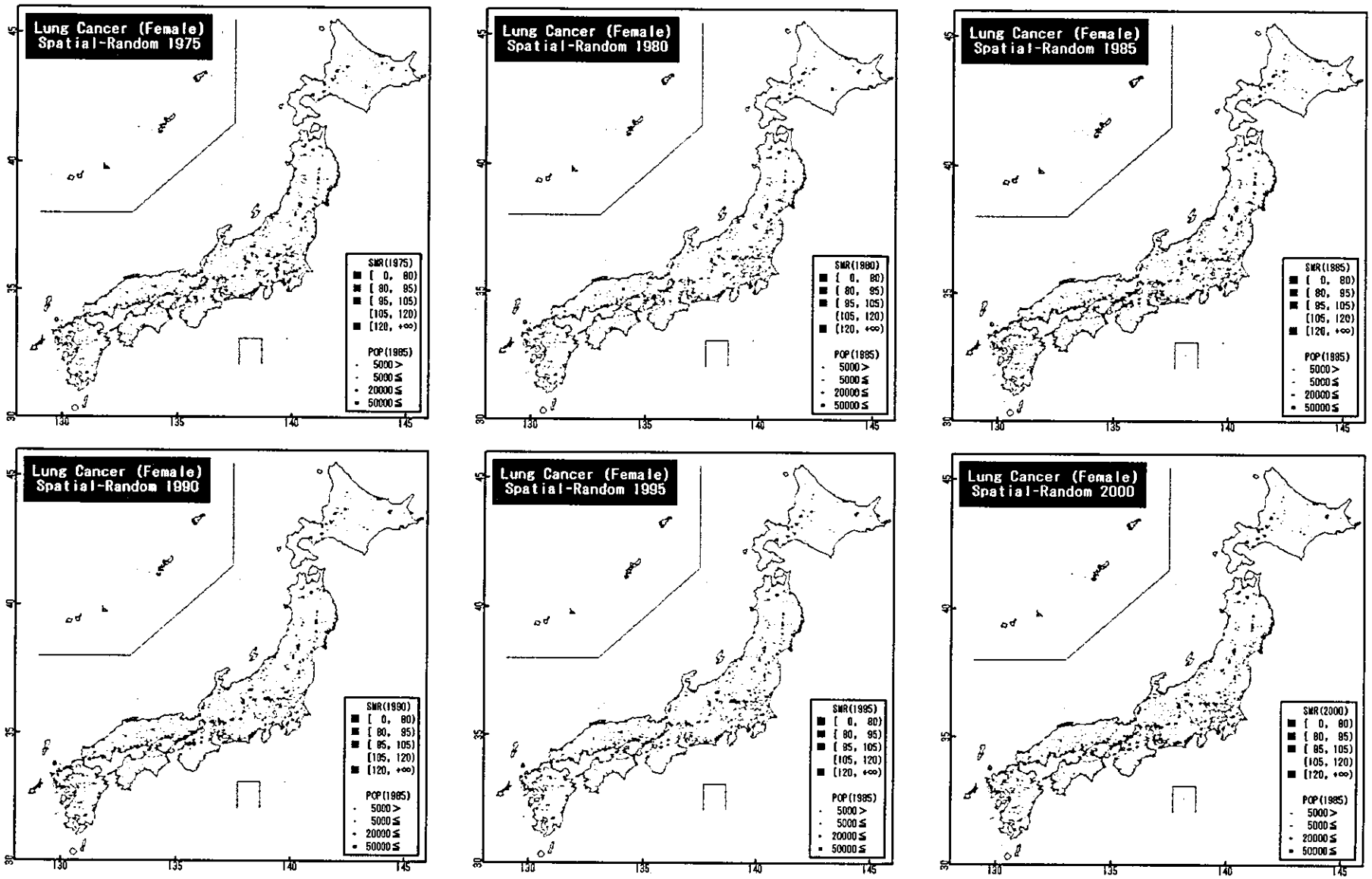


図３．女性の肺がん死亡に関するランダム効果アニメーション地図(1975年～2002年)の一部抜粋

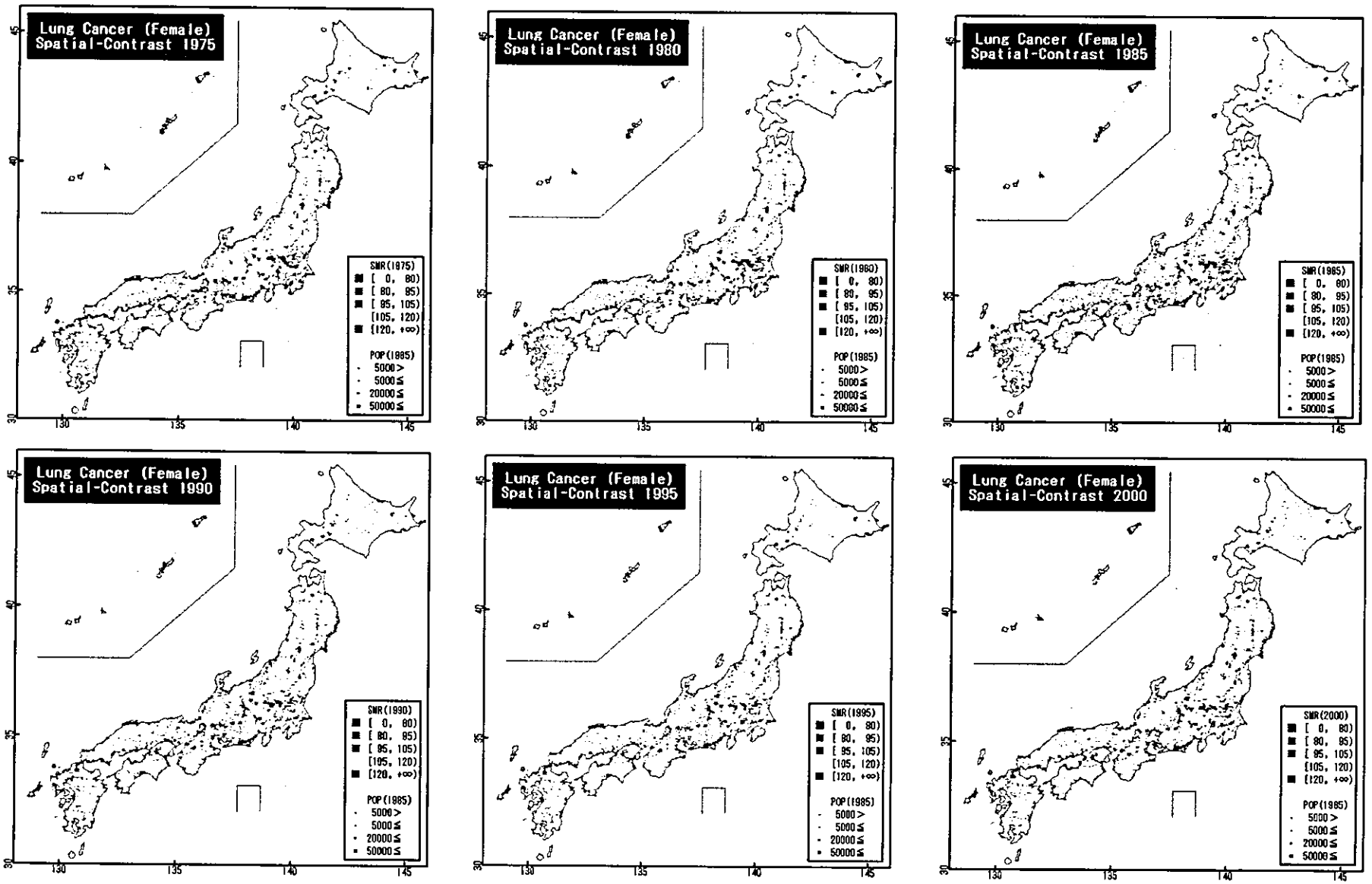


図4．女性の肺がん死亡に関する地域間コントラストSMRアニメーション地図(1975年～2002年)の一部抜粋

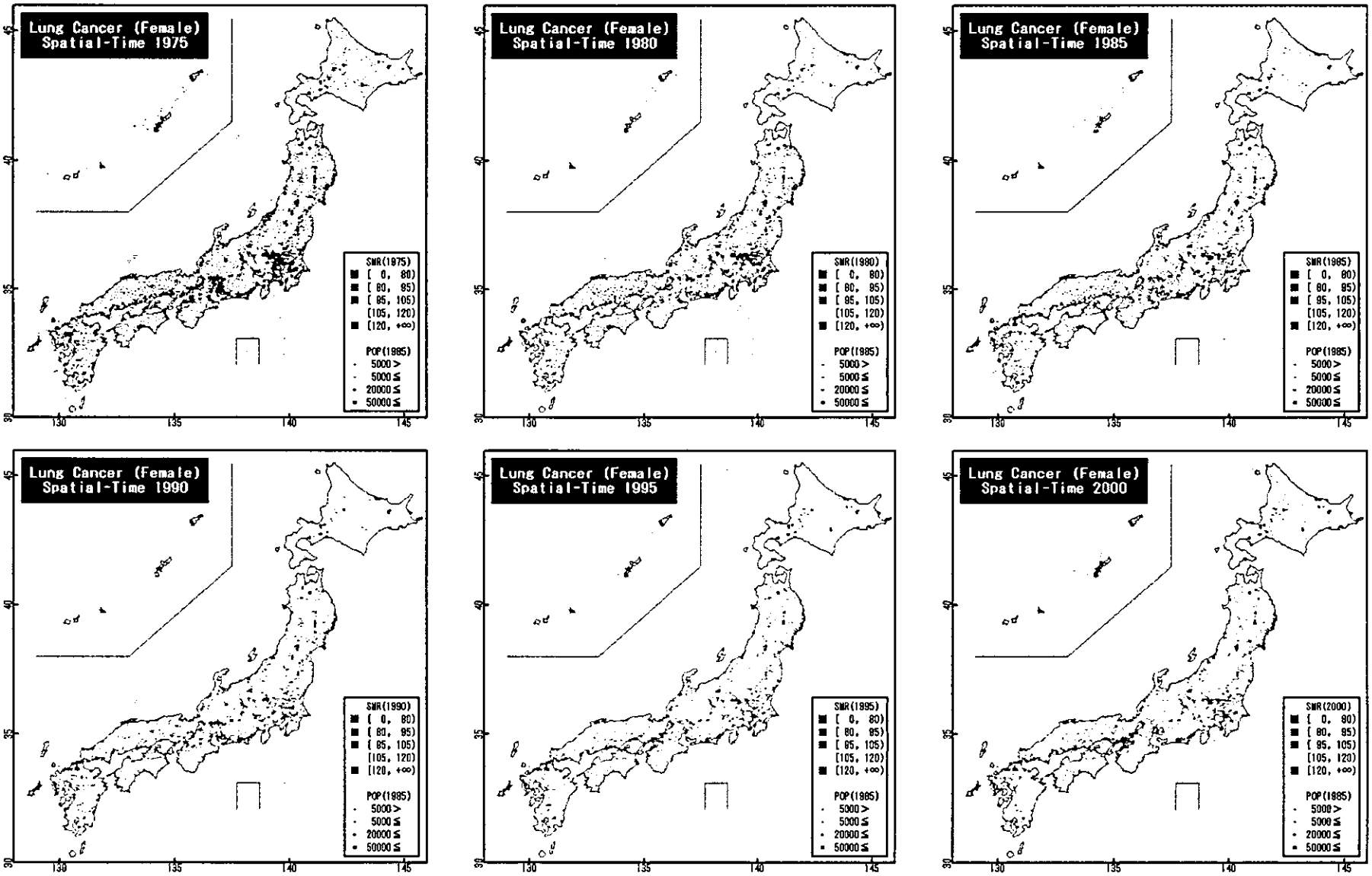


図５．女性の肺がん死亡に関する時空間最適化修正 SMR アニメーション地図(1975 年～2002 年)の一部抜粋

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 東京都老人総合研究所 副参事研究員

研究要旨

癌の罹患・死亡の動向の検討に、Age Period Model や Age Cohort Period Model がよく用いられる。肺がんは、男では 1990 年代に減少傾向がみられるが、女では 10 年程遡って減少傾向が観察される。喫煙習慣との関連では、男では、1925 年出生コホートまで、一貫して生涯喫煙率の上昇があったこと、一方、女性では 1915 年生まれ以前の出生コホートに比し、1925 年生まれまで、生涯喫煙率の減少が確認され、出生コホート効果として認識されることを報告してきた。日本の肺がんは、欧米先進国と比較したとき、中高年までは、それほどでないが、高齢者で高い位置にある。将来予測に対しては、近年の健康増進法施行があり、禁煙効果を評価するべく、 Age Period interaction の 一次、２次効果を導入しての Model の拡張を行った。

A．研究目的

日本の男性肺がんは、欧米先進国と比して、中高年までは、それほどでもないが高齢者で高いという特徴がある。近年の肺癌死亡の動向は、男では 1990 年代に減少傾向が見られ、女では、10 年程遡って、減少傾向が観察され、Age Cohort Period Model での解析から、出生コホート効果としての認識がある。

６府県大規模コホート研究資料より、男性の生涯喫煙率は、1925 年生まれまで、一貫して若い出生コホートほど増加し、そこでは 30 歳の時に 87.0% の高い喫煙率が推計されたこと、一方、女性においては 1915 年以降の出生コホートで、喫煙率の低下傾向が確認されたことを報告してきた。近年、男性の喫煙率は、徐々に低下してきたとはいえ、欧米先進国に比しては未だ高い位置にある。近年施行された健康増進法は禁煙者の割合を増加させ、喫煙者割合を低下させるものと思われる。このような状況下では、解析 Model としては、よく用いられる Age Period Model や Age Cohort Period Model に Period との interaction を考慮した Model を開発し将来予測を行い互いに比べあう必要がある。

B．研究方法

Age Cohort Period Model を用いての将来予測は、最近の若いコホートの動きを取り込みすぎると、長期予測が不安定になるという特徴がある。今回、Age Period Model に Period Age interaction を許す

(1) $\log(R_{ij}) = \alpha_i + \beta_j +$

$$+\left(Period_j - 1980\right)_+ \times \left(\delta_0 + \delta_1 \times age_i + \delta_2 \times age_i^2\right)$$

また、 Age Period Cohort Model に Period Age interaction を許す

$$(2)\ \log\left(R_{ij}\right) = \alpha_i + \beta_j + \gamma_k$$

$$+\left(Period_j - 1980\right)_+ \times \left(\delta_0 + \delta_1 \times age_i + \delta_2 \times age_i^2\right)$$

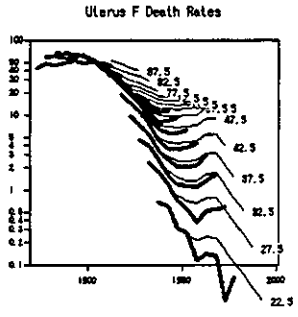
等の方法を開発した。Period Age interaction は、Period effect が 年齢によって異なる（高齢層で減少、若壮年層で増加、その逆の高齢層では増加、若年層では現象）等の場合への対応を考慮したものとなっている。

（倫理面への配慮）

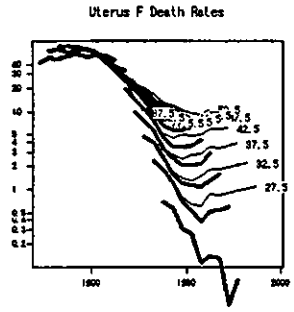
データファイルは個人識別情報を用いない。結果は統計的側面のみに使用し、生体試料は収集していない。

C．研究結果

がんの部位として、高齢層の動きと、若年層の動きで、違いが顕著な部位として子宮がん死亡の動向がある。従来の APC Model での予測は、最近の若年者の増加



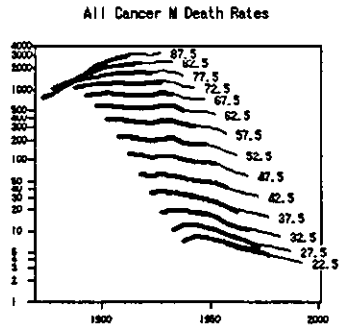
傾向はしばらく続くがその後減少し、一番若いコホートでの動きにひき連られた結果となった。死亡率が 10 万対 1 に近くなる 20 歳の後半以降のデータを用いて Period Age Interaction を許して予測は次のようである。



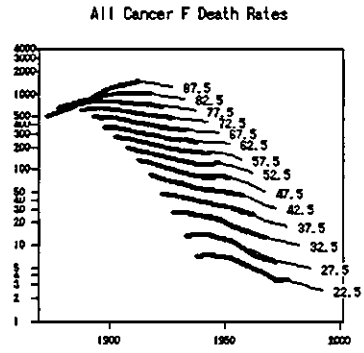
即ち、若い層での下げ止まりからの増加傾向が今後ともしばらくは続くという予測になった。

全部位のがんの結果を以下に示した。

男



女



男性の高齢層では、増加が続き、一方女性では、全年齢層でゆるやかな減少傾向が観察される結果となった。

D．E. 考察と結論

癌の罹患や死亡の動向の検討には、Age Period Model に加えて Age Cohort Period Model がよく用いられるようになった。Age Cohort Period Model では、出生 Cohort の動きが反映されるというよい面がある一方、長期予測では、若年コホートの動きに引きずられるという悩ましい面がある。

今回、 Period Age Interaction 効果を許す Model を開発し、応用を行った。Age Cohort Period Model では、例えば、女性の子宮がん死亡の動向のように、若い出生コホートでの動きに引きずられながらも、period age interaction を反映した予測が可能となった。全がん等での予測は、高齢層における男女差が浮き彫りになったこと、若年層での減少効果は、若い出生コホートほどより大きいこと等が示唆された。

肺がんの動きで言えば、喫煙習慣との関連で、戦中終戦時以降のたばこもの不足が、日本人の肺がんを国際的には、低いところに位置付けたが、戦後の長い間、喫煙率の高い時代が続いたことが、今日、日本人男性の高齢期における肺がん死亡率を、国際的に高いところに位置付けている可能性がある。Age Period Cohort Model における比較的単純な動きに加えて、Period Age Interaction を加味した解析が今後重要と推察された。

F．健康危険情報

特になし

G．研究発表

1．論文発表

1) Imamura Y, Mizuno S. Cancer statistics digest: Mortality trends of rectal cancer in Japan: 1960-2000. Jpn J Clin Oncol 2004 34:107-11.

2) Yoshimi I, Mizuno S. Mortality trends of Hematologic Neoplasms (Lymphoma, Myeloma, and Leukemia) in Japan (1960-2000): with Special Reference to Birth Cohort. Jpn J Clin Oncol 2004 34:634-7.

3) S Mizuno, H Ito, N Hamajima, A Tamakoshi, K Hirose, K Tajima. : Association between Smoking Habits and Tryptophan Hydroxylase Gene C218A Polymorphism among the Japanese Population. J Epidemiol 2004;14:94-99.

2．学会発表

1) K Iwai, S Mizuno, Y Miyasaka, T Mori. Correlation between fine particles in the environmental air and causes of diseases among inhabitants. Clean Air, London. 08-22. 2004

2) 水野正一, 秋葉澄伯：RERF LSS Report 13 にみる Radiation- Dose-Response. 第15回日本疫学会総会 2005.1.21-22 (大津)

3) 水野正一、冨田真佐子、村山隆志： 喫煙、禁煙が血清尿酸値に及ぼす影響 (縦断研究) 第38回日本通風核酸代謝学会 2005.2.3-4 (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

厚生労働省厚生科学研究費補助金（第３次対がん総合戦略研究事業）
分担研究報告書

「院内がん登録の機能強化と標準化及びがん登録事業に関わる人材育成と研修の標準化」の研究

分担研究者　金子　聰　国立がんセンターがん予防・検診研究センター情報研究部
研究協力者　今村由香　国立がんセンターがん予防・検診研究センター情報研究部

研究要旨
国立がんセンター中央病院院内がん登録を全国の院内がん登録のモデルとして、院内がん登録の最初の重要作業である腫瘍見つけ出し作業（Casefinding)を支援するシステムの内部設計から、ロジックの構築、プログラミング、実装、さらには運用までのフェーズを終了した。本システムにより、外来・入院を考慮せず登録候補症例を見つけ出すことが可能となる。また、登録漏れの把握や精度管理、作業管理にも用いることができるため、がん治療施設に本システムもしくは本システムで採用した方法論を導入することにより、各施設における院内がん登録の精度向上が期待される。今後、本システムのコンセプト・運用手順の普及、システムの一般化を図り、各施設における効率的な院内がん登録の運用を確立する必要がある。さらに、院内がん登録の運用に当たっての知識やがん登録作業に関した教材の開発、さらには指導者の育成も今後の急務である。

A. 研究目的

昨年度のがん予防等健康科学総合研究事業「がん予防対策のためのがん罹患・死亡動向の実態把握の研究」班に引き続き、国立がんセンター中央病院院内がん登録を標準化モデルとし、既存の病院情報システム（HIS）との連携をとりながら標準項目を充足させるシステム構築に関しての検討を行っている。昨年度は、特に、院内がん登録の作業で最初の重要作業である「腫瘍見つけ出しシステム（Casefinder)」の開発に関するロジックの組み立てと外部設計まで終えた。今年度は、内部設計から、ロジックの修正、プログラミング、実装、さらには運用までのフェーズを終了し、一応の成果を得た。以下、今年度の成果を報告する。

B. 研究方法

腫瘍見つけ出し（Casefinding）を有効かつ効率的に行うためには、病院情報システムと連携し、既存の情報を有効に活用することが必須である。そのためには、病院情報システムから院内がん登録への情報の受け渡し、院内がん登録での情報利用の効率化、腫瘍見つけ出しのための情報活用のロジック構築、登録業務への円滑な移行と情報活用、院内がん登録業務管理への利用等を考慮しつつ開発を進める必要がある。また、今回の開発は、国立がんセンター中央病院のみで稼働するシステムではなく、一般がん治療施設においても利用可能とする必要がある。各施設では、異なった病院情報システムや電子カルテを導入しているため、今回の開発で、汎用性を重視し、病院情報システムと直接連動しない仕組みでの情報連携を基本に開発を進めることとした。基本コンセプトとしては、以下の３点とした。

① 病院情報システムのセキュリティ確保とシステムへの負荷防止のため、院内がん登録側から病院情報システム側（データベース）に直接アクセスすることは避ける
② 異なる病院情報システムもしくは電子カルテを採用している施設においても、本研究班開発の「腫瘍見つけ出しシステム(Casefinder)が利用可能となるように、csvファイルを介したデータのやり取りを行うことを基準とする
③ 腫瘍見つけ出しシステムにより日々の作業管理を実現し、作業負荷を平準化（作業負担の平均化）させることと迅速化（登録の管理）が可能となるよう配慮する

以上の開発コンセプトにより、システムの開発と実装、さらには、国立がんセンター中央病院院内がん登録での運用を試みた。

倫理面への配慮

今回の研究は、院内がん登録システムならびにそのロジックの開発に関する研究であり、研究成果としては、システム運用やそのロジックに関する内容となる。従って、個人情報を用いて解析・研究することを目的としていない。しかし、院内がん登録という性格上、システム開発中に個人情報を扱うことになるが、これに関しては、院内がん登録を国立がんセンター中央病院の公的事業として位置づける「国立がんセンター中央病院院内がん登録関係組織規程及び院内がん登録実施規程」に従い、個人情報に関して厳格な管理のもと行われている。したがって、個人情報が院内がん登録室から外部に提供されることはなく、院内がん登録情報を用いた集計結果は、国立がんセンター事業報告等の公的資料上に公表する予定である。

C. 結果

「腫瘍見つけ出しシステム」の開発と運用の結果を報告する。以下、情報並びに作業の流れに沿って説明する。

【腫瘍見つけ出しシステムの概要（図１参照）】

① 病院情報システムから院内がん登録サーバへのデータの定時自動転送。
② 腫瘍見つけ出しシステムへのデータの取り込みと腫瘍関連情報のフィルタリング、繰り返し情報の折りたたみ。
③ 取り込み情報の情報発生日時順の提示（腫瘍見つけ出し作業支援）。
④ 登録作業への移行時の院内がん登録システムとの連動と連携：腫瘍見つけ出しシステムからの院内がん登録システムの呼び出しと、個人情報ならびに追跡情報（最終来院情報、死亡退院情報）の院内がん登録システム側への自動受け渡し。
⑤ 腫瘍見つけ出し作業から登録作業終了までの作業管理支援。
⑥ 登録終了患者、もしくは非登録対象者の新規情報発生待ちでの管理：一度作業が終わった患者（登録終了もしくは登録の対象外の判定終了）について、前回の作業終了日から６ヶ月毎に新規情報発生状況を管理、③へループ。

以下それぞれの内容について説明する。

１．病院情報システムから院内がん登録サーバへのデータの定時自動転送

（病院情報システムからのデータ転送については、今回開発した「腫瘍見つけ出しシステム」の稼働する前段階の作業であり、病院情報システムの管理・保守を担当する業者でのセットアップが必要となる）

自動受け渡しのデータは、患者基本情報と腫瘍関連診療情報の２種類である。それぞれのファイル名は、ptYYYYMMDD.csv、dataYYYYMMDD.csv（YYYYMMDD は、情報取得日）としており、ファイルとデータを情報取得日で管理する。データに含まれる情報は、情報取得日から２日前に病院情報システム上で何らかの医療情報が発生した患者についての患者基本情報と腫瘍関連診療情報が抽出されることになっている。なお、抽出情報は、国立がんセンター中央病院の場合、病院情報システム運用の開始された日からの累積情報が提供されることになっている。提供データファイルのフォーマットについては、表１，表２を参照。

腫瘍見つけ出しには、腫瘍関連情報を患者が持っているかどうかが重要になる。今回のシステムでは、腫瘍関連情報として、以下の項目を設定した。

a. 病傷名
b. 病理組織診断コード、
c. 抗癌剤の使用、
d. 放射線治療の有無、
e. シンチグラム検査の有無、
f. 外科手術の有無、
g. 内視鏡切除の有無、
h. 入退院歴

上記内容は、国立がんセンター中央病院での試験的運用を考慮しており、一般病院が利用する際は、特に同じ項目にする必要は無い。各施設で、情報システムや作業手順を考慮した内容を設定することは可能である。

表１．患者基本情報のデータフォーマット

種ptYYYYMMDD.csvファイル定義			YYYYMMDD->取得対象年月日	
列項目情報				
#	列名	最大長（byte）	意味	備考
1	患者ID	8	患者を一意に特定するID	
2	氏名	20	患者漢字氏名	
3	カナ氏名	23	患者カナ氏名	半角カタカナ
4	生年月日	10	患者の生年月日	YYYY-MM-DD
5	性別	1	患者の性別	M:男 F:女 U:Unkown O:Other
6	郵便番号	8	現住所の郵便番号	
7	現住所コード	6	現住所の住所コード	
8	現住所	80	現住所	
9	初診年月日	10	患者の初診日	YYYY-MM-DD
10	最終更新日	10	データの最終更新日	YYYY-MM-DD
11	抽出年月日	10	データを抽出した日付	YYYY-MM-DD
12	最終来院日	10	最終受診日	YYYY-MM-DD

例えば、３月 20 日に何らかの医療情報が発生した患者については、３月 22 日にそれまでの