

厚生労働科学研究費補助金（がん予防等健康科学総合研究事業）
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

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。これらの作業は、一般的に、レコードリンケージと呼ばれ、その際、もっとも重要となるのが、「同一人に関するレコードの同定」である。本研究では、個人同定処理がコンピュータによって自動化されるように、数理モデルに基づく理論の開発とそのソフトウェアの開発を行った。

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）によって経験的に判断するしかない。これらの更新作業には数ヶ月の時間を要することもあり、個人同定作業の標準化と効率化は急務である。そのためには、コンピュータによる自動化が不可欠であると考え、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行するためのソフトウェアの開発を行った。

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、一致型を得る。例えば、共通項目が（氏、名）でレコードのペアが{(佐藤、健一)、(祖父江、健一)}とすると、一致型は(X、○)または(0,1)と記述できる。すべてのペアの一致型を調べることで、一致型別のペア数を算出できる。ここで、一致型別のペア数の内訳は、同一人に関するペア数と別人に関するペア数の和であることに注目すると、一致型別のペア数は、同一人ペア数と別人ペア数の和を平均とするポアソン分布で近似できることが分かる。したがって、同一人ペアおよび別人ペアにおいて各項目が一致する確率を未知パラメータとすれば、一致型別ペア数表からこれらは推定され、その結果として、一致型別の同一人ペア数の割合が推定される。ところが、このモデルでは、別人ペアにおける各項目の一致する確率に、レコード値の「稀ぐあい」が反映されない。すなわち、別人ペアにおいて氏が（佐藤）で偶然一致する確率は低くないが、（祖父江）で一致することは極めて稀である。こ

の場合、別人ペアである可能性が極めて低くなることで、相対的に同一人ペアである可能性が高くなることが望ましい。そこで、我々は、「別人ペア数が同一人ペア数に比べて非常に多い」ことを利用して、レコード値の稀ぐあいを反映させた同一人ペア割合の推定方法を開発した。

②ソフトウェアの開発

上記①での理論が実際のデータファイルで実行できるようにソフトウェアを開発した。ソフトウェアは3つの独立したプログラムとして開発した。一つ目は、照合したい2つのデータファイル（テキスト）と共通項目を設定することで、一致型別ペア数を算出する。このとき、同時に、別人ペアにおける項目別一致頻度を調べる。二つ目は、一致型別ペア数の表をもとに、同一人ペアおよび別人ペアにおいて各項目が一致する確率を数理モデルのもとで推定する。三つ目は、同一人ペアにおいて各項目が一致する確率をもとに、同一人ペアの候補とその同一人ペア割合を提示する。プログラムは、プロトタイプとして CUI を用いて開発し、Windows OS 上で単独動作する。

（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（15万レコード）と県関連のデータ（1.5万レコード）の

個人同定処理を行った。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。パーソナルコンピュータ（CPU: Athlon1900+, RAMメモリ: 1G バイト）によってソフトウェアが要した時間は1時間弱であった。

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。一方で、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. レコードリンケージにおける個人同定処理の自動化について. 第5回博多シンポジウム. 2003年12月. 博多.

2) 佐藤健一・早川式彦・隅田治行・大瀧 慈. 関連資料間のレコードリンケージ自動化のための統計的方法. 第14回日本疫学学術総会. 2004年1月. 山形.

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

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 1997 年までの 23 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、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、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、それぞれの部位のがんの死亡危険度が、時空間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。

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、 i 市、 j 市区町村の第 t 年次での観察死亡数を Y_{ijt} とし、その帰無仮説の下での期待値をとるとき、その標準化死亡比は、

$$SMR_{ijt} = \frac{Y_{ijt}}{\hat{Q}_{jt}}$$

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 \hat{Q}_{jt} 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市区町村毎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年次別死亡数は、下記の平均を持つポアソン分布に従うことを想定した、

$$E(y_{it}) = \hat{\alpha}_t e^{\bar{I}_{it} + \hat{\alpha}_{it}};$$

$$i = 1; \dots; n; t = 1; \dots; k;$$

ここで、 \bar{I}_{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、 $\hat{\alpha}_{it}$ 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、の事前分布は、平均 1、分散 $\hat{\alpha}_t$ 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 N_t において集計した死亡数 $\sum_{i \in N_t} Y_{it}$ および期待死亡数 $\sum_{i \in N_t} \hat{\alpha}_{it}$ も使用する。以下、 Y_{it} や $\hat{\alpha}_{it}$ と区別して使用するために、それぞれ $Y_{it}[t]$ および $\hat{\alpha}_{it}[t]$ 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]

Step 1. 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、

$$SMR_{\hat{e}t} = \frac{\sum_{i=1}^n Y_{it}}{\sum_{i=1}^n \hat{\alpha}_{it}}; t = 1; \dots; k;$$

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

$$w_{ij}(r) = \begin{cases} \frac{1}{q} \frac{a}{1 + \cos \frac{A_{ij}}{r} x_{ij}} & ; \quad 0 \leq x_{ij} \leq r; \\ 0 & ; \quad x_{ij} > r; \end{cases}$$

により定める。ただし、 q は重み*i*に関する総和が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 3. 次式で定義される CV 型規準量を $Q(M)$ 最小とする人口数 M を求める、

$$Q(M) = \sum_{i=1}^n \sum_{j=1}^n \frac{w_{ij}(r_i(M)) \log \frac{1+y_{ij}}{1+\hat{q}_{ij}}}{\sum_{j=1}^n w_{ij}(r_i(M))} \bar{A} \log \frac{1+y_{ij}}{1+\hat{q}_{ij}} ;$$

Step 4. 前 Step で最適化された M 値を \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対数変換値の重み付き空間平均を次式により算出する、

$$\hat{\mu}_{[t]} = \sum_{j=1}^n w_{ij}(r_i(\hat{M})) \log \frac{1+y_{j[t]}}{1+\hat{q}_{[t]}} ;$$

$$i = 1; \dots; n; t = 1; \dots; k;$$

Step 5. 市区町村毎の $\hat{\mu}_{[t]}$ を年次*t* について、ノンパラメトリック平滑化を行い、その結果を、 $f_{[t]}(i) = 1; \dots; k$ と記す。

Step 6. 年次*t* 毎に、変動効果の分散母数 σ^2 に関する尤度関数

$$L(\hat{\mu}_{[t]} | \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{y_{i[t]} A(\hat{\sigma}^2)}} \exp \left\{ -\frac{1}{2\sigma^2} \frac{(y_{i[t]} - \hat{\mu}_{[t]})^2}{A(\hat{\sigma}^2)} \right\} ;$$

を定め、その値を最大化する最尤推定値 $\hat{\sigma}_{[t]}^2$ を求める。

Step 7. $\hat{\mu}_{[t]}$ を年次*t* についてノンパラメトリック平滑化を行い、その結果を、 $f_{[t]}(i) = 1; \dots; k$ と記す。

Step 8. 各年次別に市区町村毎に、変動効果母数 $\alpha_{[t]}$ の事後平均値を次式により推定する、

$$\hat{\alpha}_{[t]} = \log \frac{y_{i[t]} + \hat{\sigma}_{[t]}^2}{\hat{q}_{[t]} e^{\hat{\mu}_{[t]} + \hat{\sigma}_{[t]}^2}} ;$$

Step 9. 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、ここで、

$$RR_{i[t]} = e^{\hat{\mu}_{[t]} + \hat{\alpha}_{[t]}} = e^{\hat{\mu}_{[t]}} \frac{SMR_{i[t]} + \hat{\alpha}_{[t]}}{e^{\hat{\mu}_{[t]} + \hat{\alpha}_{[t]}}}$$

である。人口数の大きい都市の場合、 $\hat{\alpha}_{[t]} = \hat{\mu}_{[t]} - \hat{q}_{[t]}$ 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、 $\hat{\alpha}_{[t]}$ の値が大きくなるために、 $RR_{i[t]}$ の値はその町村の近傍に関する時空間平滑値 $SMR_{i[t]}$ に近くなる。最後に、年次変動に関してより高い精度を持たせるために、市区町村毎のに対して、 $SMR_{i[t]}$ に関するノンパラメトリック平滑化を施し、その結果をと記し、下記の式による調整を行う、

$$SMR_{i[t]} = SMR_{i[t]} \exp \left\{ \frac{1}{n} \sum_{j=1}^n \hat{q}_{[t]} \log RR_{j[t]} \right\} ;$$

[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 $SMR_{i|t}$ を5個のカテゴリに層別化して色を対応させ、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)

本研究では、個人を識別できるような情報および生体試料を扱っていない。

C. 研究結果

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かになった。

D. 結論

ポアソン・ガンマモデルに基づく経験ベイズ法とノンパラメトリック平滑化の適用による年次毎の性別市区町村別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ことが確認できた。

E. 健康危険情報

特になし

F. 研究発表

1. 論文発表

1) Yanagihara, H., Ohtaki, M.: Knot-Placement to avoid over fitting in B-spline scedastic smoothing, Communications in Statistics, Simulation and Computation, 32(3), 771-785, 2003.

2. 口頭発表

1) Ohtaki, M.: Spatial-Time Distribution of Major Cancer Mortalities in Japan -- A statistical analysis of municipality-specific demographic data. International Conference on Cancer Research Prevention and Treatment, Silchal (India) 2003.12

2) 大瀧 慈, 川崎 裕美, 佐藤 健一, 中山 晃志, 柳原 宏和, 山口 直人: 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, 第14回日本疫学会学術総会, 山形. 2004.1

G. 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

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 Cohort Period Model が用いられるようになった。肺がんに関しては、近年の減少傾向にコホート効果が認識されている。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。このことの背景を知る目的で、過去の大規模コホート研究をもとに、生涯喫煙率を検討した。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確認された。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た。40歳時点での喫煙率が11・12%から6.7%へとほぼ半減していたことが確認された。

A. 研究目的

近年の肺癌死亡の動向をみると、男では1990年代に減少傾向がみられ、女では、10年程遡って、減少傾向が観察される。癌の罹患や死亡の動向の検討に、Age Cohort Period Model での検討から、出生コホート果たす効果が大きいことが認識されている。

肺がんの動向に関しても、出生コホート効果が大きく、今回、そのコホート効果がどの時点で形成されたのか背景を探るべく、過去の大規模コホート研究をもとに生涯喫煙率の検討を行った。

B. 研究方法

曾田、平山らによって計画実施された265,000人を対象とした6府県大規模コホート研究資料を用いた。1965年基礎調査資料中、現在喫煙状況（非喫煙、前喫煙、不規則喫煙、毎日喫煙）喫煙経験（喫煙開始年齢、喫煙期間の年数）をもとに1965年

までの生涯喫煙状況を再構築した。

（倫理面への配慮）

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

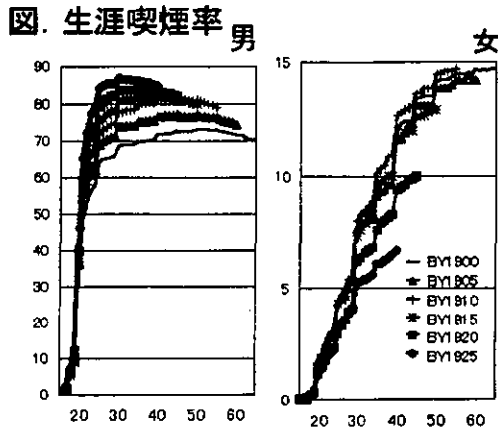
C. 研究結果

5年ごとの出生年での対象者数は、出生年1900年（男3042人、女3106人）1905年（3615, 3980）、1910年（4336, 4864）、1915年（4236, 5070）、1920年（4526, 5885）1925年（4587, 5062）であった。図に、対象者の生涯喫煙率を示した。

男性では、一貫して若いコホートほど増加の傾向が見られた。1925年生まれにおいては、30歳の時に87.0%の高い喫煙率が推計された。

女性においては1915年以降の、若年の出生コホートで、喫煙率の低下傾向が確認さ

れた。



1925 年生まれにおいては 30 歳で 5.2%、40 歳で 6.7% と低い値に留まった。1915 年生まれ以前の cohorts では 40 歳での喫煙率は 11-12% であった。

D. 考察

男性では 20 歳から 30 歳にかけて喫煙率がピークに達している。今回の対象者は、1965 年次 40 歳以上の人たちである。喫煙開始年齢の多くが 20 歳前後ということを見ると、この cohort 研究対象者の喫煙習慣の形成は多くが 1945 年終戦時以前に形成されているものと解釈される。日本人男性に対して何時の時期に喫煙習慣が形成され、1925 年出生 cohort では 87% という値になったが十分検討されなければならない。男性肺がんの出生 cohort 解析からは、この出生 cohort までの肺がん死亡率の上昇が確認されているので、この点に関しては今回の結果は符合している。

その後の出生 cohort で、若いときの喫煙率が減少したかということが男性の肺がん死亡率の近年の動きを理解する上で重要であるが、一職域 cohort での解析結果を発表し、その後の出生 cohort で一旦喫煙

率の減少を確認している。

女性における今回の結果は重要と考えたい。即ち、男性が喫煙率を上昇させている間、女性は喫煙率を減少させていて、この間の 10 年間ということが、女性の肺がん死亡率が男性に 10 年ほど前だって減少していることを考える上で有用と思われる。

肺がんの地理分布に関しては、地域としての特徴があり、喫煙習慣は大きいものであるが、他の Risk-factor との interaction 等を十分に考える必要がある。今回の結果を踏まえ、喫煙率における地域差を十分に検討し将来の発展へとつなげる必要がある。

E. 結論

男性では喫煙開始年齢との関係で、この cohort 研究対象者の喫煙習慣の形成は多くが終戦時前に形成されたものと解釈される。一方、終戦時以降のたばこ不足からの解消が日本人の喫煙率を高いところに引き上げたことも推察される。一方、女性では、喫煙開始年齢が 20-40 歳と年齢の幅が広い中にあり、1915 年以降の出生 cohort において喫煙開始を順次遅らせ、たばこ不足の時代に生涯喫煙率を全体的に低いところに抑制した効果があったと結論される。その後の cohort の動向を加味する必要はあるが、男女間におけるこの十年間の違いが、肺癌動向の出生 cohort 効果に影響を与え、近年の肺がん死亡率の減少が女性では、男性に先んじて観察される一要因と結論したい。また、高齢期における日本女性の長寿の要因との関連で興味深い。

今回の結果から、一つの結論として、日本における肺がん死亡率の動向を考える上で、若い時の喫煙動向の把握が重要である

ことが導かれる。

F. 健康危険情報
特になし

G. 研究発表

1. 論文発表

1) Kaneko S, Ishikawa B, Yoshimi I, Marugame T, Hamashima C, Kamo K, Mizuno S, Sobue T. Projection of lung cancer mortality in Japan. Cancer Sci 2003; 94: 919-923.

2) Marugame T, Mizuno S. Mortality trend of lung cancer in Japan 1960-2000. Jpn J Clin Oncol 2003 33(3) 148-9

2. 学会発表

1) Mizuno. Association Between Smoking Habits and Tryptophan Hydroxylase Gene

Polymorphisms in Japanese. 第26回日本がん疫学研究会 2003年6月. 札幌

2) 水野正一. 大規模職域男性コホートにみられた喫煙習慣の出生コホート差. 第62回日本癌学会総会. 2003年9月. 名古屋

3) 水野正一. 6府県大規模コホート研究にみられた生涯喫煙率の出生コホート効果. 第62回日本公衆衛生学会総会 2003年10月. 京都

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

院内がん登録の機能強化と標準化およびがん登録事業に係わる人材の育成と研修の標準化

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

地域がん登録の登録精度を向上させるために必須である院内がん登録の全国的な整備に関して、国立がんセンター中央病院の院内がん登録を標準化のモデルとし、検討を行った。登録工程については、米国の院内がん登録を参考に、1) がん患者の見つけ出し(casefinding)、2) 一時ファイルへの保管、3) 数ヶ月後(4~6ヶ月)の登録対象判定、4) 登録情報の抽出、5) 登録情報のコード化、6) 予後調査(follow-up)、7) 集計・解析・報告書作成の7工程を標準とした。さらに、最初の3工程(がん患者の見つけ出しから登録対象判定まで)について病院情報システムとの連携を考慮したシステムを検討した。患者見つけ出しについては、医師による報告形式は採用せず、院内情報システムとの連携により自動的に行うシステムを検討した。登録すべきか否かの判断と登録情報の抽出についても、病院情報システムからの自動的に移送するシステムを作り上げること検討したが、院内がん登録に必要なすべての情報が病院情報システムに含まれていないこと、さらにそのロジックが確立していないことから、当面、手作業により行うこととした。今後、今回検討したロジックでの患者見つけ出しを実際に行い、問題点の把握と解決、さらには、登録情報の抽出と登録を円滑に行うための支援プログラムを開発する必要がある。

A. 研究目的

地域がん登録の登録精度を向上させるためには、院内がん登録の整備が必須であるが、国立がんセンターにおいてですら、院内がん登録の整備が進んでいない。院内がん登録標準化に向けての動きは、地域がん診療拠点病院の指定要件に院内がん登録の実施が含まれたこと、厚生労働科学研究費補助金効果的医療技術確立推進事業「がん診療の質の向上に資する院内がん登録システムの在り方及び普及に関する研究」班(主任研究者:山口直人)において「地域がん診療拠点病院院内がん登録標準項目とその定義」を策定し、普及に努めていることなど徐々に伸展しつつあるが、標準化を促進するためのモデル施設が存在しない。

そこで、本研究では、国立がんセンター中央病院の院内がん登録を標準化に関するモデルとし、既存の病院情報システム(HIS)との連携をはかりながら標準項目を充足させるシステムを構築することを第一の目的とする。また、開発・運用を通じて蓄積される知識・経験・システムをがん登録関係者の教育・研修用に利用できる形で要約することも目的の一つである(第2目標)。最終的には、院内がん登録を整

備し、地域がん登録の精度が向上することが最終の目標となる。

今年度は、第一目標である国立がんセンターにおける院内がん登録の標準化モデルの検討を行った。

B. 研究方法

1) 院内がん登録の運用に関する検討

米国で運用されている院内がん登録の作業手順を参考とし、我が国における標準的院内がん登録の作業手順に関する検討を行った。

2) HISとの連携に関する検討

国立がんセンター中央病院は、電子カルテ化はなされていないものの、巨大なHISが整備されている。今回HISと院内がん登録の連携に関しての検討を行った。

3) 院内がん登録構築に向けての検討

院内がん登録は個人情報を含めた診療情報が登録されるため、HISに準じた安全対策を施す必要がある。安全対策を検討するとともに、作業の効率化に関する検討を行った。

1)、2)、3)ともに将来的な地域がん診療拠点病院への展開を考慮し、国立がんセンター以

外の施設でも利用可能となるよう検討に配慮を行った。

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される。

C. 結果

1) 院内がん登録の運用に関する検討

米国における院内がん登録の作業工程を参考にした登録工程手順は、以下の通りである。

- a. がん患者の見つけ出し (casefinding)
- b. 一時ファイル (suspense file) に保管
- c. 数ヶ月後 (4~6 ヶ月) に登録すべきか？
既登録か？の確認
(多重がんの判定)
- d. 登録情報の抽出 (abstracting)
- e. 登録情報のコード化 (coding)
- f. 予後調査 (follow-up)
- g. 集計・解析、報告書作成

この工程手順は、がん登録の起点を医師からの報告とするのではなく、“院内がん登録”が積極的に行う患者の見つけ出しが起点となる。従って、医師への負担が軽減され、また、系統的ながん患者の見つけ出しが可能となる。

2) HIS との連携に関する検討

上記の作業を HIS との連携により、どの程度自動化できるかについての検討を行った。

a. がん患者の見つけ出しについての検討

HIS からの情報の受け渡しは、日単位で行い、病院において何かしらの診療行為がある 1 日になされた患者情報を HIS が開設以来の履歴情報付きで受け取る仕組みを検討した。患者見つけ出しに関して HIS から受け取る情報

は、“がん”の診断に関連する以下の情報のみとした：

- ① 医事課の病傷名
- ② 病理組織診断名
- ③ 放射線治療歴
- ④ 化学療法歴 (詳しくは、化学療法プロトコール名)
- ⑤ 手術歴

さらに①から⑤の情報が発生した日を受け取る。国立がんセンター中央病院の場合、約 1500 名/日のがん患者情報、約 9 万件を HIS から受け取ることになる。これらすべてについて、確認を行うことは不可能であるため、件数を減らし、自動的に登録すべき症例を抽出する方法を検討した。その上で、人的作業により登録すべきかどうかの判断を行う。自動化を図る上で、設定した条件は以下の通りである。

- ・ reference date (登録開始日) の設定
院内がん登録開始日を 2004 年 4 月 1 日に設定し、この日以前に診断・治療情報が発生している患者については、院内がん登録開始以前の症例として、登録対象外とすることとした。ただし、同じ化学療法プロトコールで登録開始日以降の加療がなされているような場合を除くために、同一情報については、開始日と終了日と頻度を残し、他の情報は削除 (折りたたむ) することにより情報量を削減することとした。つまり、A というプロトコールの化学療法がこれまで 10 回なされている場合、HIS から受け取る場合の情報は、10 回の治療分とその施行日であるが、折りたたみの後は、A の治療が開始された日付と最終日、さらには 10 回の頻度のみの 1 情報に集約される。この A 治療開始日が、院内がん登録開始日以前である場合は、過去の診断例として登録対象としない。しかし、2004 年 4 月 1 日以降、B 治療が開始された場合は、登録候補対象として、登録すべきかの確認を行う。
- ・ がん関連以外の情報の削除
病理組織情報は、すべての標本の情報が提供されるため悪性の組織診断以外の情報も含まれる。従って、悪性診断以外の情報を削除する。ただし、脳腫瘍に関しては、良性であっても残すこととする。削除対象は、SNOMED コードで登録されているため、概ね良性と良性・悪性境界を示す病理組織コード 5 桁目で判断する。ただし、国際疾病分類腫瘍学第 3 版にお

ける変更を反映させるため、変換マスターも作成し、登録漏れの内容に配慮する。

b. 一時ファイル (suspense file) への保管・管理についての検討

a 患者見つけ出しで見つけられた患者について、すぐさま登録すべきかどうかを判断し、登録情報を診療録から抽出するとその後に発生した診断・治療情報を再度確認する必要が発生し、作業量が倍増する。従って、米国の方法にならい、患者見つけ出しからある一定の期間をおいた後に、登録すべきかの判断と登録情報の抽出を行うこととした。これらの状態を把握するために、a の作業において HIS から提供された患者情報に対して、その患者の現在の状況を指し示す“ステータス”を付記し、管理することとした。

ステータスの遷移	遷移の理由	遷移するステータス	発生データの発生状況
0: 初期値	患者のデータが登録されていないこと	0: 初期値	発生しない
1: 判定対象	4ヶ月経過	1: 判定対象	発生しない
2: 判定対象外	がん検出	2: 判定対象外	発生しない
3: がん検出	確定作業開始	3: がん検出	発生しない
4: 調査継続	確定作業継続	4: 調査継続	発生しない
5: がん確定実施中	がん確定実施中	5: がん確定実施中	発生しない
6: がん確定保留	がん確定保留	6: がん確定保留	発生しない
7: がん確定	がん確定	7: がん確定	発生しない
8: チェック済	チェック済	8: チェック済	発生しない

図1. ステータスについて:あるステータス(左端)の状態の患者に新たな状況が発生すると真ん中の列のステータスに移行する。

ステータスの遷移フロー

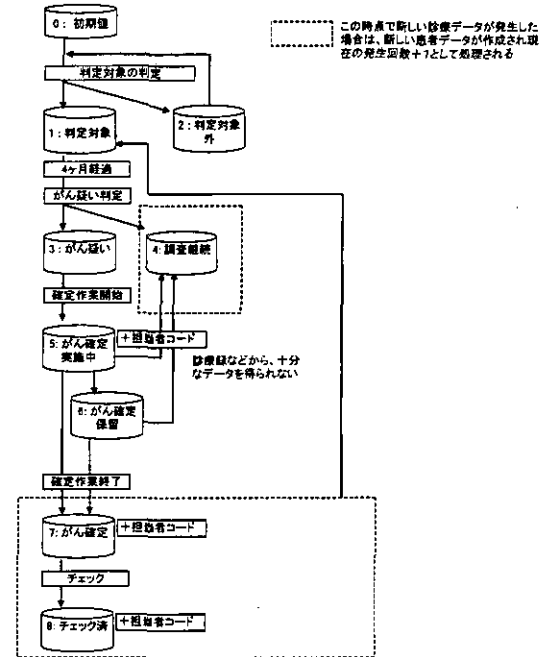


図2. ステータスの変遷フロー

ステータスが一巡し、もとの状態の戻った段階で次の新しい情報がきた場合、発生回数(N)を1から2にあげて再度同じ動きを繰り返す。このようなスパイラルな状況が新たな情報が来るたびに続く事になる。

c. 数ヶ月後(4~6ヶ月)に登録すべきか? 既登録か?の確認(多重がんの判定)に関する作業の検討

この作業は、現時点では、判断ロジックが組み立てられていないため、自動判定のプログラムが組めないため、HIS から得られた情報を総合して、人的作業により判断する事とした。ただし、判定の判断支援については、コンピュータによりプログラムを組んで行う。判定判断支援プログラムのインターフェースは以下のようにする予定である。

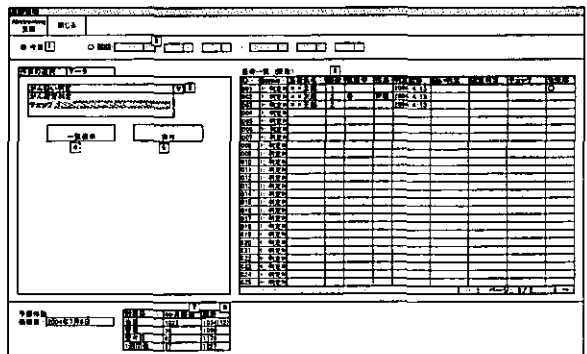


図3. 登録対象者の管理画面のイメージ:左の四角3の選択で、患者見つけ出しで見つかり、登録作業に移る

べきかどうかの患者一覧、登録作業に移行した患者管理のための一覧、抽出作業が終了し、ダブルチェック待ちの患者一覧が選択できる。

d. 登録情報の抽出 (abstracting) 並びに e. 登録情報のコード化 (coding) に関する検討

情報の抽出が最も電子化された情報を利用する価値のある部分である。しかし、登録に必要な情報が現時点ではすべて HIS に含まれていないこと、抽出ロジックが検討されない段階で、システムへの組み込みは、戻り作業や繰り返しが多く発生し、開発時間と費用を無駄にする可能性が高いとの判断から、現段階では、診療録を実際に人手で抽出し登録する事とした。実際の登録作業ならびにコーディング作業については、既に地域がん診療拠点病院向けに開発・配布されている院内がん登録システムを応用することにより、作業負担の軽減を図ることとした。

f. 追跡調査ならびに、g. 集計・解析、報告書作成に関する検討

追跡調査については、既に国立がんセンター中央病院ではシステム化され、病院事業として運用されているため、この部分の検討は現時点では、行っていない。しかし、HIS から受け取る情報に、受診時点での住所情報も盛り込まれるため、HIS から受け取る際に少なくとも、

- a. その時点での生存が確認される
- b. その時点での最新住所が更新できる

の2点は確認できる。従って、院内がん登録システムへのこれら2点の情報を毎回更新するようプログラムを作成することで検討を進めた。

3) 院内がん登録構築に向けての検討

国立がんセンター中央病院の院内がん登録の実際の運用は、別棟のがん予防・検診研究センターで行い、診療録閲覧は、診療録を管理している中央病院調査課で行うこととした。当初、1) カルテ閲覧室において登録票に転記し、入力作業を行う案、2) カルテ閲覧室にノート型コンピュータを携行し、入力の後、院内がん登録サーバに転送する2案が検討されたが、紛失・置き忘れ等による個人情報ファイルの漏出が危惧されたため、病院とがん予防・検診研究センターを結ぶ閉じたネットワークを用いて登録作業を行うこととなった。ネットワークへの接続は、がん登録専用の端末のみ可能となる

構造となっている。

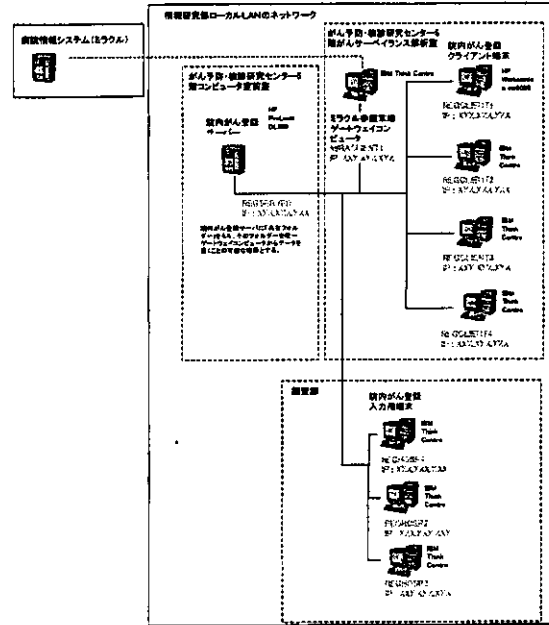


図3. 院内がん登録の閉じたネットワーク

また、HIS との接続に関しては、HIS に負荷や障害を与えることなくすることを考慮し、院内がん登録のネットワーク側からは、アクセスできない仕組みにし、HIS に接続する1台の端末コンピュータ（ゲートウェイコンピュータ）にがん患者見つけ出しデータを HIS 側から置き、院内がん登録サーバ側が取りに行く仕組みとした。

登録すべきかどうかの判断の際には、参照専用の HIS を使い、院内がん登録側からは、オーダーや変更等が不可能である状況での利用とした。

D. 考察

院内がん登録において、もっとも重要で最初のステップとなる工程が患者見つけ出し (casefinding) である。これまで、多くの施設においてこの作業は、主治医に一任されてきた。主治医が院内がん登録側に患者に関する報告をなすという患者見つけ出し方法は、受け身的 (passive) であり、これまでの経験や米国での結論から報告漏れが多く、施設を受診する全がん患者を把握するには問題があるとされている。また、主治医による報告の場合、入院単位とならざるをえず、入院期間短縮傾向が進み、複数の入院治療が増えることを考慮すると、重複報告が発生し、効率が悪い。また、外来での診断・治療患者の把握も難しくなる。従って、院内がん登録がある程度独立して患者見つけ出し作業が行える環境の整備することが必要となる。今回の国立がんセンター中央病

院の院内がん登録では、医師による作業負担を極力削減し、がん登録側で患者見つけ出し作業を行う設計とした。

国立がんセンターの場合、新規入院患者が年間4500例程度であり、外来のみでの診断・治療例も増えていることから、年間5000から6000例のがん登録対象者が発生することが予測されている。しかも、疑い患者についても、診療録を参照し登録するかどうかを判断することになるため、実際のところ、患者見つけ出しで対象となる患者は、年間で1万例近くになるのではないかと予想される。このような作業を手動で行うことは困難であるため、今回自動的に患者見つけ出しをする手順の検討とプログラムの設計を行ったが、実際の運用は、次年度以降になる。今後、実働に伴い、修正が必要となると考えられる。患者見つけ出しのシステムについては、ある程度安定した運用が可能となった段階で、地域がん診療拠点病院向けに変換し配布することも視野に入れている。

患者見つけ出し以後、登録情報を抽出しコード化して、登録する必要があるが、医師による登録に頼る場合、院内がん登録の登録項目に関して、医師から理解を得る必要があり、標準化されたがん登録を維持することは難しい。逆に、医師以外、たとえば診療情報管理士による登録となると、情報の信頼性が低くなるとの危惧が予想される。今回の国立がんセンターでの院内がん登録では、これまで我が国には存在しない、“腫瘍登録士”を教育し、この腫瘍登録士を中心に登録作業を行うことを計画している。その教育資料を作成するに当たっても、今回の院内がん登録構築は重要な情報源となる。また、腫瘍登録士によるがん登録を支援するコンピュータアプリケーションについても、必須であるため、来年度以降、検討・開発を進める必要がある。特に病期分類に関しては、各学会・研究会が定め、外科系の医師を中心に広く使用されている癌取扱規規約や、国際的に広く用いられているUICC TNM分類、また地域がん登録で標準として用いられている臨床進行度分類があり、これらを統合した形で登録・変換できるシステムが必要である。米国においても複数の病期分類システムががん登録で稼働していた（SEER summary stage, Extension of disease, AJCC (TNM) 分類の3つ）が、2004年より統合した登録形式に変更されている。この登録形式では、病期分類で用いられる部位により異なる基本情報をそれぞれ登録し、コンピュータプログラムにより各病期分類に自動変換するシステ

ムである。我が国においても同様のシステムを構築すべく、来年度の研究において、このシステムを開発し、運用段階まで終了させる予定である。

E. 結論

院内がん登録の作業工程に関する検討とがん登録において最初の作業となる患者見つけ出しのシステムについて検討・開発を行った。今後、実際の運用により、完成度の高いシステムとし、一般病院向けに変更する予定である。また、登録支援プログラムの開発も進める必要がある。

F. 健康危険情報

特になし

G. 研究発表

1. 論文発表

- 1) 金子 聰、祖父江友孝: 肺癌の今後. 最新医学 58:2462-2469, 2003
- 2) Kaneko S, Ishikawa KB, Yoshimi I, Marugame T, Hamashima C, Kamo K, Mizuno S, Sobue T: Projection of lung cancer mortality in Japan. *Cancer Sci* 94:919-923, 2003
- 3) Yoshimi I, Kaneko S: Mortality trend in female breast cancer in Japan: 1960-2000. *Jpn J Clin Oncol* 33:254-255, 2003
- 4) Kaneko S, Sobue T: Mortality trends of all cancer sites in Japan: 1960-2000. *Jpn J Clin Oncol* 33:51-52, 2003
- 5) Kaneko S, Tamakoshi A, Ohno Y, Mizoue T, Yoshimura T: Menstrual and reproductive factors and the mortality risk of gastric cancer in Japanese menopausal females. *Cancer Causes Control* 14:53-59, 2003
- 6) Kaneko S, Yoshimi I: Mortality trend of stomach cancer in Japan: 1960-2000. *Jpn J Clin Oncol* 33:105-106, 2003
- 7) Mizoue T, Yoshimura T, Tokui N, Hoshiyama Y, Yatsuya H, Sakata K, Kondo T, Kikuchi S, Toyoshima H, Hayakawa N, Tamakoshi A, Ohno Y, Fujino Y, Kaneko S: Prospective study of screening for stomach cancer in Japan. *Int J Cancer* 106:103-107, 2003

2. 学会発表

- 1) 金子 聰、水野正一、丸亀知美、吉見逸郎、濱島ちさと、祖父江友孝、がん死亡の将来予測に関する検討. 第62回日本癌学会総

会 名古屋 2003. 9. 26

- 2) 金子 聰. 地域がん診療拠点病院における院内がん登録の整備. 第 12 回地域がん登録全国協議会 福井 2003. 9. 12
- 3) 加茂憲一、金子 聰、吉見逸郎、丸亀知美、今村由香、濱島ちさと、祖父江友孝、味木和喜子、津熊秀明、水野正一. がん罹患数の全国推定方法に関する検討. 第 14 回日本疫学会学術総会 山形 2004. 1. 23
- 4) 吉見逸郎、水野正一、丸亀知美、加茂憲一、今村由香、金子 聰、祖父江友孝. 悪性新生物による死亡数の将来推計の試み. 第 14 回日本疫学会学術総会 山形 2004. 1. 23
- 5) 今村由香、金子 聰、加茂憲一、祖父江友孝. 院内がん登録標準化に向けての取扱い規約と TNM 分類についてのがん登録担当者への支援の在り方について：主要 5 部位における病気分類からの考察. 第 14 回日本疫学会学術総会 山形 2004. 1. 23

II. 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

Ⅲ. 研究成果の刊行に関する一覧表

別紙 5

研究成果の刊行に関する一覧表

論文

発表者氏名	論文 タイトル名	発表雑誌名	巻号	ページ	出版 年
Yoshimi I, Ohshima A, Ajiki W, Tsukuma H, <u>Sobue T.</u>	A Comparison of Trends in the Incidence Rate of Lung Cancer by Histological Type in the Osaka Cancer Registry, Japan and in the Surveillance, Epidemiology and End Results Program, USA.	Jpn J Clin Oncol	33	98-104	2003
Kaneko S, Ishikawa K, Yoshimi I, Marugame T, Hamashima C, Kamo K, Mizuno S and <u>Sobue T.</u>	Projection of lung cancer mortality in Japan.	Cancer Sci	94	919-923	2003
Kaneko S, <u>Sobue T.</u>	Mortality Trend of all Cancer Sites in Japan: 1960-2000.	Jpn J Clin Oncol	33	51-52	2003
Yoshimi I, <u>Sobue T.</u>	Mortality trend of liver cancer in Japan: 1960-2000.	Jpn J Clin Oncol	33	202-203	2003
Marugame T, <u>Sobue T.</u>	Mortality trend of pancreas cancer in Japan: 1960-2000.	Jpn J Clin Oncol	33	416-417	2003
Marugame T, <u>Sobue T.</u>	Mortality trend of ovarian cancer in Japan: 1960-2000.	Jpn J Clin Oncol	33	595-596	2003
<u>Sobue T.</u>	An Epidemiological Aspect of Lung Cancer -Increase in mortality and anti-smoking measures-	JMAJ	46	521-524	2003

研究成果の刊行に関する一覧表

論文

発表者氏名	論文 タイトル名	発表雑誌名	巻号	ページ	出版 年
<u>Ajiki W</u> , Tsukuma H, Oshima A.	Survival rates of childhood cancer patients in Osaka, Japan.	Jpn J Clin Oncol	34	50-54	2004
<u>Ajiki W</u> , Tsukuma H, Oshima A.	Cancer incidence and incidence rates in Japan in 1998: estimates based on data from 12 population-based cancer registries.	Jpn J Clin Oncol	33	241-245	2003
Verdecchia A, Mariotto A, Gatta G, Bustamante-Teixeira MT, <u>Ajiki W</u> .	Comparison of stomach cancer incidence and survival in four continents.	Eur J Cancer	39	1603-1609	2003
Nomura E, Tsukuma H, <u>Ajiki W</u> , Oshima A.	Population-based study of relationship between hospital surgical volume and 5-year survival of stomach cancer patients in Osaka, Japan.	Cancer Sci	94	998-1002	2003
Ioka A, Tsukuma H, <u>Ajiki W</u> , Oshima A.	Ovarian cancer incidence and survival by histologic type in Osaka, Japan.	Cancer Sci	94	292-296	2003
Ioka A, Tsukuma H, <u>Ajiki W</u> , Oshima A.	Trends of uterine cancer incidence in Japan in 1975-1998.	Jpn J Clin Oncol	33	645-646	2003

研究成果の刊行に関する一覧表

論文

発表者氏名	論文 タイトル名	発表雑誌名	巻号	ページ	出版 年
石黒信吾, <u>味木和喜子</u> , 上堂文也, 他.	「食道癌と他臓器重複癌 EMR 時代を迎えて」 疫学 および病理学の立場からみ た食道癌と他臓器重複癌.	胃と腸	38	283-290	2003
森川康英, 細井創, <u>味 木和喜子</u> , 津熊秀明, 加藤忠明.	「横紋筋肉腫」 横紋筋肉 腫の疫学.	小児外科	35	8-12	2003
津熊秀明, <u>味木和喜子</u> .	「大腸癌の診断と治療 最 新の研究動向」 大腸癌の 疫学的研究 疫学 国際比 較.	日本臨床	61	25-30	2003
本間正巳、高橋孝、國 井一彦、横山紘一、松 <u>田徹</u> 、深井正仁、佐藤 幸雄	山形県がん登録（第 59 報） 平成 11 年（1999 年）標準集 計.	山形県病医誌	37(2)	160-178	2003
<u>松田徹</u>	山形県がん登録最近 15 年間 の生存率と死因調査.	山形県医師会報	619	44-46	2003
横山紘一、 <u>松田徹</u> 、佐 藤幸雄	山形県の胃癌の地域別比較.	山形県医師会報	620	13-14	2003
<u>陶山昭彦</u>	がん検診と個人情報保護.	泌尿器外科	16(9)	908-921	2003
<u>陶山昭彦</u> ら	前立腺がん検診における個 人情報保護とインフォーム ド・コンセント.	泌尿器外科	16(9)	925-958	2003