様式 C-19

科学研究費補助金研究成果報告書

平成２３年　５　月　３１　日現在

機関番号：１６４０１
研究種目：若手研究（B）
研究期間：　2009　～　2010
課題番号：21790499
研究課題名（和文）大規模医療情報データベースを用いて疾患発症リスクを予測する統計学的モデルの構築
研究課題名（英文）Study of the statistics model to be predictable of disease risk by using large-scale medical information data base
研究代表者
中島 典昭　(NORIAKI NAKAJIMA)
高知大学・教育研究部医療学系・助教
研究者番号：00335928

研究成果の概要（和文）：
大規模医療情報データを用いた病態推移予測モデルの構築を目指し、高知大学医学部附属病院の医療情報データベースの検査値時系列を用いて、個人における検査値の変遷を予測するモデルの構築を行った。糖化ヘモグロビンの検査値時系列データを対象として集団の従う分布を背景に個々のデータを評価しながら解析できる潜在曲線モデルを集団の特徴抽出と個々のデータ予測を可能なモデルとして検討した。さらに採用したモデルの適応範囲を調査する為に、数値シミュレーションによって得られた理想的な模擬データを用いてモデルの適応範囲を明らかにした。

研究成果の概要（英文）：
The purpose of this research is construction of the condition transition forecasting model that uses the large-scale medical information data. The latent curve model to be predictable of transition of the inspection value in each patient was constructed by using the inspection value time series data in the medical information data base at Kochi University hospital. In the latent curve model, it is possible to analyze it while evaluating individual data in the background of distribution that the group follows it. The potential curve model was examined as a model by whom the individual patient data was able to be forecast. The range of the adjustment of the model was clarified by using the ideal simulated data that had been obtained by the numerical simulation.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,400,000	420,000	1,820,000
2010 年度	700,000	210,000	910,000
年度			
年度			
年度			
総　計	2,100,000	630,000	2,730,000

研究分野：　医歯薬学
科研費の分科・細目：境界医学・医療社会学
キーワード：医療情報システム、医療データベース

１．研究開始当初の背景

近年の高度な高齢化社会において膨大な医療費は大きな問題となっている。これに対処すべく疾病予防や健康維持には関連省庁や地方自治体において積極的な取り組みが

なされているところであるが、疾病予防のための方策をたてる上では十分な情報の確保が必須となる。一方、ほとんどの病院において医療情報は病院情報システム内でデータベース化されており、各患者に対する診療における一次利用に加えて、統合的な二次利用も可能な段階に入っているといえる。全国に散在するこれら医療情報を統合し利用することで、疾病予防方策を検討するための情報量は十分に確保できるものと考えられる。

病院によっては施設の特徴や地域的特性などで蓄積されている医療情報データに疾患の偏りがあり（栗原等、2004)、疾患によってはその症例数が僅少であり十分な統計を得られない場合がある。しかし、多数の医療施設の医療情報データを収集し、統合データベースを構築することでその問題を克服することが出来ると考えられる。つまりデータの質的な差異に対して十分な検証を行いながら量的な増加を図ることが可能なのである。このようにして得られた大規模医療情報データベースには様々な年齢、疾病の個人が投薬歴、手術歴と共に検査値時系列として記録されており、これらを詳細にわたり統計モデル化することで任意の状態にある個人に対して様々な角度からの疾病リスク予測が可能となると考えられる。

本研究に先立って、高知大学医学部附属病院の統合医療情報システム（IMSI：奥原等、2003）に内包されている医療情報データベース中の検査値時系列データを用いて、病態推移を行う予測モデルの構築を行った。検査値時系列の動向予測においては対象となるデータ量が不十分である場合、その予測の不確定さは避け難い問題となる。そこで検査値時系列の動向が既に確認されている多数の症例を背景として、潜在曲線モデルの考え方を基にして時系列データを特徴づける変数の事前分布を求めた。求めた事前分布を用いて予測対象の時系列データの動向予測を行う病態推移予測モデルを提案した（渡部等、2007)。

２．研究の目的

近い将来実現されると予想される統合医療情報データベースより得られる大規模医療情報データを解析して様々な疾病予防のための病態推移予測モデルの構築を目指している。

本研究では高知大学医学部で所有する医療情報データを利用して大規模医療情報データを扱う統計学的手法を開発し、更に病態推移予測モデルを開発する。このデータベースは開院以来約 30 年分の患者基本情報、病名歴、薬歴、手術歴、検査歴等が蓄積されていおり、長期に渡る患者の状態が把握できるであろう検査値データが抽出できる。特に疾患の診断基準となっている検査値の時系列データに注目し、発症予測などができるモデルを構築し、モデルの適応範囲を明らかにする。

３．研究の方法

個人における検査値時系列データの動向を確率過程に従う現象として捉え、それのモデル化を行った。

解析に用いた潜在曲線モデルは、時間 t_i に測定された検査値 x_i で構成される要素 $\mathbf{X}_i=(x_i,t_i)$ から成る検査値時系列データ $\mathbf{X}=(\mathbf{X}_1,\ldots,\mathbf{X}_K)$ が、線形的な時間発展とその周りでの揺らぎとして記述できる場合、それを特徴づける変数 θ によってどの程度的確に検査値時系列データが表現されているかは、尤度関数を用いて評価できる：

$$P(\mathbf{X}\mid\theta)=\prod_{i=1}^{K}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(a+bt_i-x_i)^2}{2\sigma^2}\right]$$

ここで $\boldsymbol{\theta}=(a,b,\sigma)$ と定義した。変数の事前分布 $P(\boldsymbol{\theta})$ を導入することによって事後分布を得る。

$$P(\theta\mid\mathbf{X})=\frac{P(\mathbf{X}\mid\theta)P(\theta)}{\int P(\mathbf{X}\mid\theta')P(\theta')d\theta'}$$

これにより、検査値（病態）の推移は条件付

$$P(X_{K+1}\mid\mathbf{X})=\int P(X_{K+1}\mid\theta)P(\theta\mid\mathbf{X})d\theta$$

き確率で予測される。

この方法によりデータ量が少ない検査値時系列の時間発展の予測を安定して行うことが可能となる。

生活習慣病の診断基準に利用されている検査について、特に糖尿病の診断指標の一つである糖化ヘモグロビン(HbA1C)を高知大学医学部附属病院の医療情報データベースから検査値時系列データを抽出し、開発した病態推移予測モデルを適用してモデルの変数分布を求めた。これらモデルの変数分布は集団の要素である個人の既往歴などに影響されるため、集団の適切なカテゴリー分けが必要となり、そのカテゴリー分けを明確にしながら得られた変数分布を用いて病態推移予測を行った。変数の事前分布の設定には解析者の主観介入を極力排除する必要があるため、検査値推移が十分に判っている多人数(N人)の検査値時系列データ $\{\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(N)}\}$ を

背景にして変数の“事前分布”を求めることとした（渡部等、2007）。事前分布 $P(\boldsymbol{\theta})$ を N 人の検査値時系列データから求められる事後分布 $P(\theta|\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(N)})$ に置き換える。事後分布 $P(\theta|\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(N)})$ を求めるには分布を特徴づける超変数 $\boldsymbol{\omega}$ を導入し、N 人の検査値時系列データの集団としての特徴を抽出する必要がある。

$$P(\boldsymbol{\theta}|\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(N)}) =$$
$$\int P(\boldsymbol{\theta}|\boldsymbol{\omega})\left\{\prod_{i=1}^{N}\int P(\mathbf{X}^{(i)}|\boldsymbol{\theta}')P(\boldsymbol{\theta}'|\boldsymbol{\omega})d\boldsymbol{\theta}'\right\}$$
$$P(\boldsymbol{\omega})d\boldsymbol{\omega}/P(\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(N)})$$

ここで超変数 $\boldsymbol{\omega}$ の積分を一般的な数値積分の手法で行うには困難を伴う。そのためマルコフ連鎖モンテカルロ法（MCMC）による超変数のサンプリングを行い、得られたサンプルから変数の事後分布を求める。サンプリングにおいて超変数 $\boldsymbol{\omega}$ から $\boldsymbol{\omega}^{*}$ へ移行する際のアクセプタンスレイトはメトロポリスヘイスティング法に従い定義する。
変数 $\boldsymbol{\theta}=(a,b,\sigma)$ について、各変数間には相関がないものと仮定して変数の分布を独立な形で定義した。

次にモデルの適応範囲を明確にする為に、シミュレーションで作成した疑似データをつかって作成した時の変数分布のパラメーターを推定した。このとき、患者数と一人まあたりのデータ数をそれぞれ変化させ、従来の元となるパラメーターを直接計算する方法と比較を行いながら、モデルの有効性と適応範囲を確認した。

上記のモデル構築のち複数の検査値から統合的に病態を予測できるモデルへの拡張を目指した。ここでは検査値の多くは独立した動態を示さず、相関して変化すると考え、集団が従う事前分布に検査値に相関があるような形でのモデル（たとえば多次元正規分布）を用いて表現した。

４．研究成果

本課題では糖尿病の診断基準の一つである HbA1c の検査値時系列データを対象として集団の従う分布を背景に個々のデータを評価しながら解析できる潜在曲線モデルを集団の特徴抽出と個々のデータ予測を可能なモデルとして検討した。HbA1c の検査値時系列データは性別でのカテゴリー分け行った。その際、糖尿病等の病名歴のある患者や、インシュリンなどの糖尿病用薬の薬歴がある患者は除いた。また解析では１人あたり３点以上のデータが存在する必要がある為、その制限も同時に適応した。結果、患者数は 102 人でデータ数は 369 点（１人あたり 3.6 点）が解析対象として抽出された。データそれぞれカテゴリーにおいて直線近似をもとにした潜在曲線モデルを適応した。分布関数の違いによる比較とパラメータ分布の決定を行い先行研究と整合性のある結果となった。しかしながら、HbA1c は過去においては必ずしも必要な検査値ではなく、測定されていること自体何かしら糖代謝の異常が疑われている状態であることは明らかな為、正常な状態から異常状態（罹患）へ変化する仮定を推測したい本研究においては、解析対象として偏りがあるデータであり十分ではないことが推測された。今後は病院データではなく検診データなど利用しモデルの精度を高める必要があると考える。

さらに採用した潜在曲線モデルの適応範囲を調査する為に、数値シミュレーションによって得られた理想的な模擬データを用いてモデルの適応範囲を調査した。個人のデータが従う直線のパラメーター分布（事前分布）が分かっているとして擬似的に検査値時系列データを発生させ、その疑似データに対して潜在曲線モデルを用いて元のパラメーターを推定する方法を用いた。個人の傾向から集団の特徴を導出する従来型の解析手法(Naive Model)との比較からサンプル数が比較的少ない場合でも集団の特徴を正確に推測できることが示された。たとえば、１患者あたりのデータ数（K）、患者数(N)で事前分

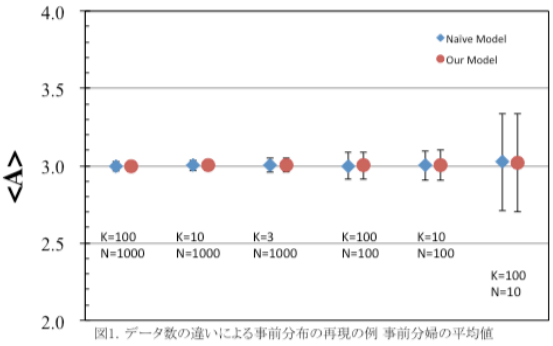

図1. データ数の違いによる事前分布の再現の例 事前分婦の平均値

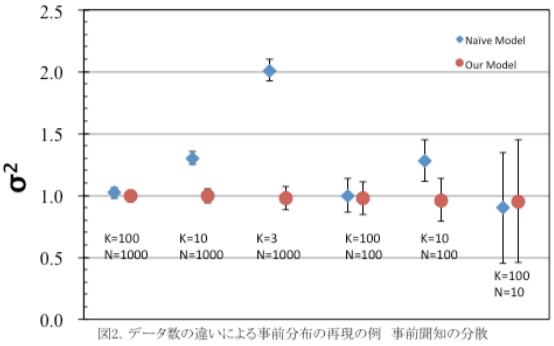

図2. データ数の違いによる事前分布の再現の例 事前開知の分散

布のパラメーターである平均値〈A〉と分散σ^2を我々のモデルで推定した。1000 セット行った場合の平均値と標準偏差の比較を〈A〉を図１に σ^2を図２にそれぞれ示す。

平均値の推定では、十分なデータ数を確保できれは、元のパラメーターの値を精度よく推定でき、従来型の解析手法とも同様にデータ数に起因する誤差の範囲でもとのパラメーター値を推定できている。一方、分差の値では、従来型の解析手法では、十分にデータ数がないと推定することができていないが、我々のモデルでは、図２より解析対象とする一人当たりのデータ数がある程度少ない状況であっても、患者数を十分に大きくすることによって、もとのとなる分布を純分再現できることが分かった。

これらの解析を通して、検討した潜在曲線モデルには、揺らぎの取り扱いをより厳密にする必要があり、個人に起因する検査値のゆらぎと測定誤差などの外的な要因によるゆらぎを個別に取り扱うことが重要であると考えられた。したがって当初のモデル自体を変更せざるおえなくなり、補助期間中には、計画していたように課題を十分に達成することができず成果発表に至ることができなかった。しかしながら、研究の基盤とデータについては十分に整備することができたので、今後早急に問題を解決し、モデル構築を達成でると期待している。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

なし

６．研究組織
(1)研究代表者
中島 典昭（NORIAKI NAKAJIMA）
高知大学・教育研究部医療学系・助教
研究者番号：00335928