【第77回生涯教育講座】

医療業務基幹系へのデータマイニングの適用

津本周作[1)]　　平野章二[1)]
津本優子[2)]

キーワード：データマイニング，病院情報システム，多重スケールマッチング，
ラフクラスタリング，一般化線形モデル

要　旨

Rapid progress in information technology has enabled us to　store all the information in a hospital information system, including management data, patient records, discharge summary and laboratory data. Although the reuse of those data has not started, it has been expected that the stored data will contribute to analysis of hospital management. In this paper, data from several university hospitals were analyzed. The results show several interesting results, which suggests that the reuse of stored data will give a powerful tool to support a long-period management of a university hospital.

１．はじめに

診療情報の電子化が1980年代からはじめられて以来ほぼ20年が経過し，病院の医事会計情報から検査データを始めとした診療部門のデータ，さらに数年前から大きく進んだ診療録の電子化に至るまで，ほとんどの診療データ，特に，テキスト・数値形式のデータはデータベースとして蓄積されるようになってきた[1)]。電子カルテの実装とそのカルテの検索効率の向上が果たされれば，電子カルテによる「ゆりかごから墓場まで」なる個人データベースの蓄積も実現が近づきつつある。検査データや医事会計情報に関しては，OLAP 的な手法によって，一部が二次利用されてきているが，データマイニングあるいは古典的な統計的手法等を含めた高度な二次利用についての議論はまだはじまったばかりである[2,3)]。我々人間のデータ解析がどちらかといえば，症例に対する深く短期的な視野での考察が特徴的であるのに対し，計算機によるデータ解析で最も特徴的なことは「横断的解析」によって，違った視野を獲得できる点である。二次利用によって新たな視点で獲得された知識を利用し，大学病院の特性をマクロでとらえ，いかなる経営と診療が将来望まれるかについ

Shusaku TSUMOTO et al.
1）島根大学医学部医学科医療情報学講座
2）島根大学医学部看護学科基礎看護学講座
連絡先：〒693-8501　出雲市塩冶町89－1

ての指針を得るためのツールとなることが期待される。

我々はこれまで病院情報システムの稼働から蓄積されてきた長期のデータから，いかなる情報が抽出でき，将来の病院経営および診療の支援に役立てることができるについて，様々な視点から検討してきた[4,5,6]。

病院経営に関する解析では，千葉大学附属病院のデータを用いた結果ではあるが，大学病院の特色として，疾患の頻度をみると，特殊な感染症や悪性新生物を含めた専門的な治療を要する疾患が疾患の上位を占める一方，病院の収入を決める第一因子が在院日数であることが明らかとなった。さらに，疾患によって，収益性が大きく異なることが明らかとなったが，疾患によっては，入院日数が正規分布に従わず，対数正規分布をとるものが多く見られ，従来の平均在院日数は，収益性を表す指標としては不適切であることがわかった[4]。

診療支援という面からでは，本データの一部である肝炎データが，科研費特定領域研究「データマイニング」(平成13年度～平成16年度）で共通データと提供され，データマイニングがいかに診療支援に寄与できるかについての検討が進められた[7]。

我々は本科研費研究班として，類型化を通じて時系列医療データの特徴を視覚的に表現，ユーザの知識の発見を促すシステムの構築を目指し，中核となる系列の比較法および類型化法として多重スケールマッチングとラフクラスタリングを取り入れた時系列の比較分類法を開発した[5,6]。これらの方法により，GPT の推移パターンの特徴，血小板数と肝炎進行度の関係など，新たな知見を獲得し報告した。本論文ではこれらの解析結果も報告する。

なお，これらの解析は平野を除く，著者が千葉大学医学部に在籍したこと，千葉大学医学部附属病院が1980年代より検査データを蓄積しはじめた関係から，比較的大規模なデータが蓄積されてきたという経緯から千葉大学のデータを使用した結果ではある。ただし，解析手法については，データの様式さえそろっていれば，他大学，他施設においても適用可能である。島根大学医学部附属病院についても，平成18年11月に診療情報の電子化がほぼ完了し，今後同様の手法の適用も考えている。

2．収益についての解析

2.1 基幹系データベースからのデータ抽出

解析当時，千葉大学病院では基幹系データベースからデータを抽出して蓄積するデータウェアハウスは現在実装されていなかった。このため，分析に必要なデータは，患者 ID と入院日をキーにして退院時要約と患者基本情報のシステムより，抽出用のプログラムを作成して抽出した。保険点数は患者 ID 別月別に保存されているため，患者１人１入院あたりの総点数を計算したのち，患者 ID と入院日をキーにして退院時要約のデータと結合させた。

退院時要約のデータベースから抽出したデータは，1978年度～1999年度の21年分で157,618件となり，保険点数を加えたデータは，1997年度～1999年度の３年分で20,164件となった。本論文では，この20,164件の解析について示す。

2.2 解析手法

退院時要約のデータは，患者基本情報（性別・年齢・職業）・転帰・入院日数・疾患の内訳，治療法についてのデータが含まれており，これに保

険点数のデータも加えて入院費用とそれに関連する項目について，記述統計，探索的データ解析，統計学的検定，多変量解析（回帰分析，一般化線形モデル，対応分析，クラスタ分析）の手法を用いて解析した。分析にはSAS（Solaris版，Ver 6.0），SPSS 11.0J for Windows, R2.0を用いた。

2.3 解析結果

2.3.1 入院日数の分布

全症例による入院日数の分布をみると（図１），右に長く尾をひいてかなり歪んでおり，対数をとると正規分布に近くなることがわかった（図２）。この傾向は，感染症をのぞき，ほとんどの疾患で見られ，新生物全体，さらに新生物の下位に属する悪性腫瘍（肺，胃，肝）においても同様の傾向が見られた。これらの結果は，在院日数が対数正規分布をとることを予想させる。

2.3.2 入院費用に関する分布

１患者１入院あたりの保険点数の総計を入院費用とし，その分布について検討した。保険点数の総計はセキュリティ上の理由で用いず，各症例の総点数を全症例の中央値で割った値を中央値指数として表した。平均値ではなく中央値を用いたのは，保険点数の分布が正規分布に従っていなかったためである。図３に中央値指数の分布を，図４に対数変換した指数の分布をヒストグラムで示した。図に示されたごとく，入院費用の分布も入院日数と同様に対数正規分布を示している。この傾向は新生物等の下位階層の分類においても同様であった。

次節で示すように，入院費用と入院日数とには強い相関関係があり，これらの対数正規分布に近い疾患は，入院日数を反映していると考えられる。

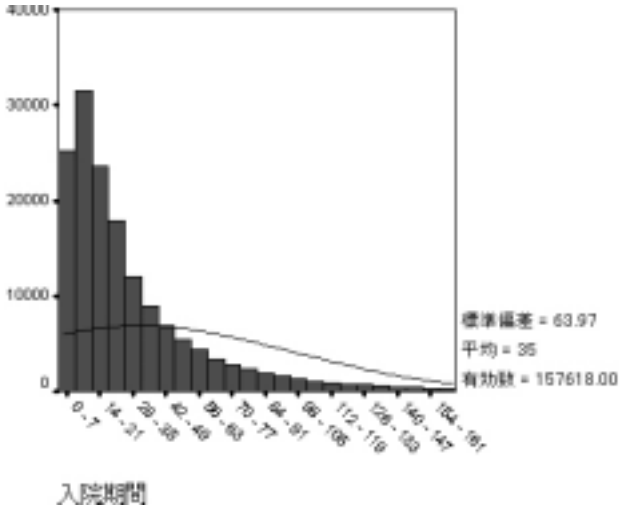

図１　全症例による入院日数の分布

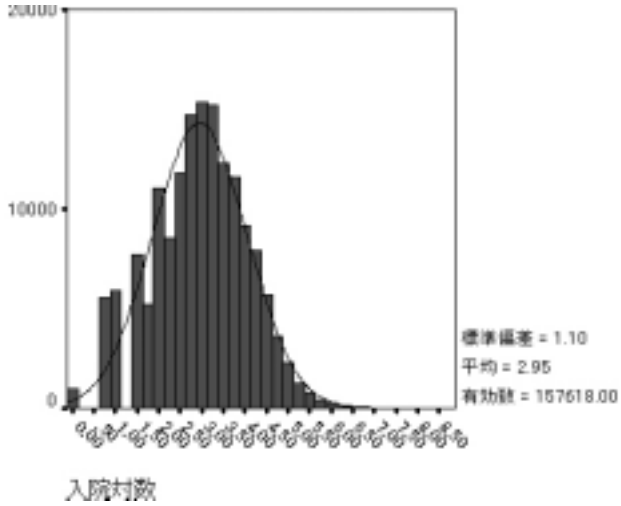

図２　対数変換した入院日数の分布

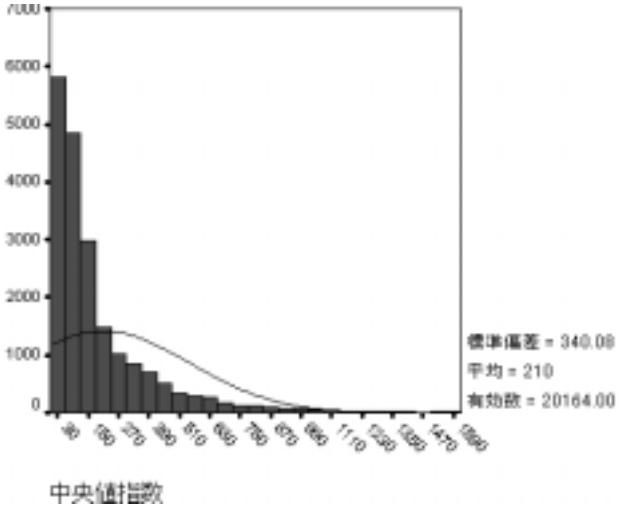

図３　入院費用に関する分布

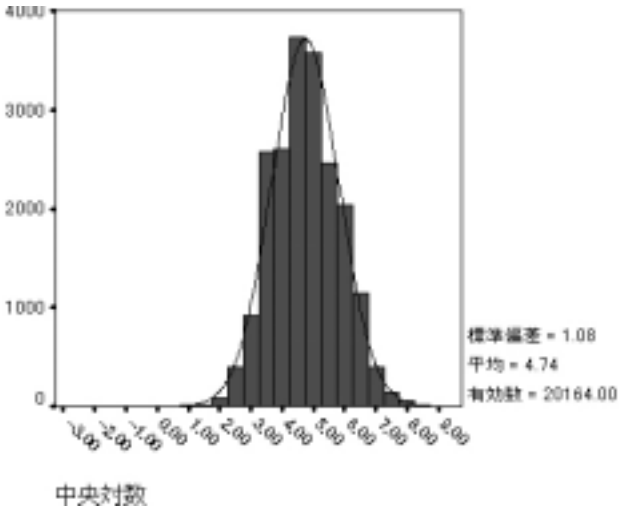

図4　対数変換した入院費用の分布

2.3.3 対数正規分布についての考察

記述統計から得られた分布の性質は，これらの分布が対数正規分布をとることを支持しているが，より客観的には，他の分布と fitting に関して定量的な比較が必要であり，今後詳細な検討を報告する予定である。

たとえ対数正規分布をとらないにしても，このような歪度の高い分布をなすことの意義は重要である。従来，病院を評価する指標として，平均在院日数が使われてきたが，平均値は，分布が正規分布あるいはそれに近似できる分布でなければその頑健性が低いことが知られており，中央値の方がより頑健性が高いことが知られている。特に，平均在院日数は，その定義から長期在院日数が必要な疾患の罹患者のみを切り捨てることで，容易に改善させることができ，数字的な操作のみの改善で，医療の質の向上の改善に関係しないことがありうる。これに対し，中央値はその病院の体制が全体として最適化されない限り，値が改善されず，長期療養者数に関しても一定量，許容されうるという意味で，より良い指標といえる。

2.3.4 入院日数と費用との相関・単回帰分析

2.3.4.1 相関分析

対数変換した入院日数と対数変換した入院費用(中央値指数）との相関係数に関して，全症例，新生物，および症例数の多い悪性新生物3種の場合について，表1に相関係数の値を示した。表に示したように，手術を施行した症例の方が手術を施行しなかった症例よりも相関係数の値が低い。これは，分布の記述統計から考察すれば，手術施行例の方が分散が高いことに由来すると考えられる。特に，これは肝・肝内胆管の悪性新生物に関して，著名な傾向を示している。

実際に，症例全体および肺の悪性新生物の手術有および手術無しの場合での，この二つの変量の散布図を図5－7に示した。

2.3.4.2 単回帰分析

相関分析で得られた結果をより詳細に検討するために，会計点数を目的変量，入院日数を説明変量として単回帰分析を行った。各点数は，前節同様，ICD 9 コード3桁の病名毎に分けて分析をおこなった（表2）。表に示すごとく，高い R^2 乗値

表1　対数変換した入院日数と入院費用との相関

	全体	手術あり	手術なし
全症例	0.837	0.829	0.779
新生物	0.867	0.844	0.826
肺・気管支の悪性新生物	0.838	0.648	0.903
胃の悪性新生物	0.827	0.738	0.801
肝・肝内胆管の悪性新生物	0.711	0.577	0.755

表2　対数変換した入院日数の入院費用に対する単回帰分析の結果

	R^2	β	定数	標準誤差
全症例	0.701	0.854	2.405	0.004
新生物	0.752	0.913	2.278	0.007
肺・気管支の悪性新生物	0.702	0.823	2.581	0.024
胃の悪性新生物	0.683	0.994	2.084	0.036
肝・肝内胆管の悪性新生物	0.505	0.875	2.407	0.047

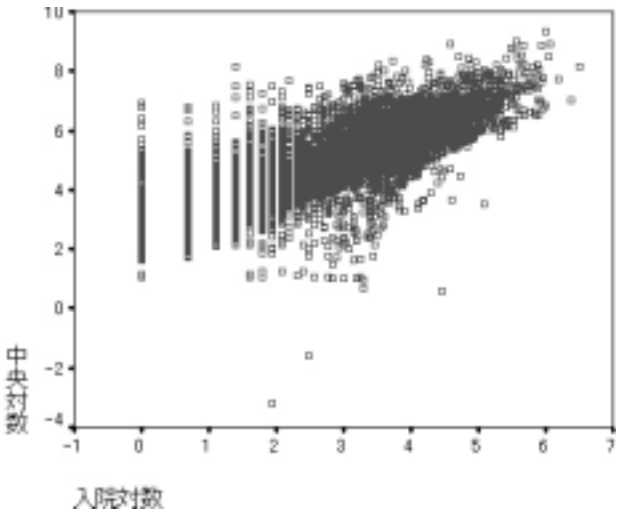

図５　対数変換した入院日数と入院費用との散布図
（全症例，横軸:入院日数，縦軸:入院費用）

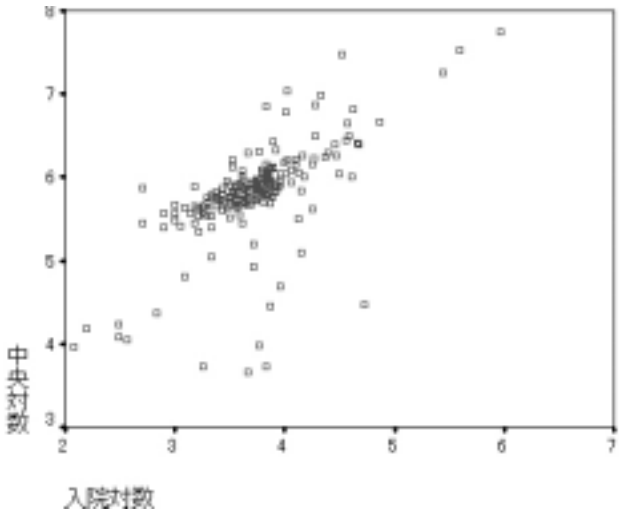

図６　対数変換した入院日数と入院費用との散布図
（肺の悪性新生物:手術あり，横軸:入院日数，縦軸:入院費用）

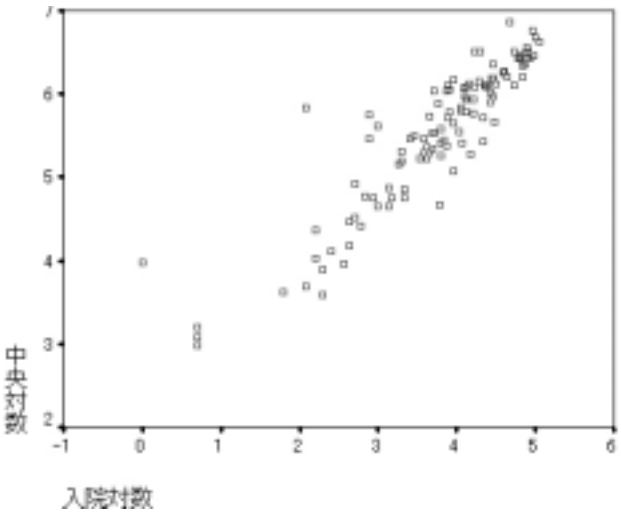

図７　対数変換した入院日数と入院費用との散布図
（肺の悪性新生物:手術なし，横軸:入院日数，縦軸:入院費用）

が得られ，入院日数が目的変量を説明するに比較的十分な変量であることが示された。また，表から説明変量の係数が各疾患の点数の特性，各疾患での収益性をとらえていると考えられた。

これらの結果は対数変換していることから，入院日数と入院費用とに関しては，べき乗のモデルがあると推定できる。例えば，日数をｘ，費用をｐ，定数をｃとすれば，

$$\ln p = \beta \ln x + c$$

より，

$$p = e^{c} x^{\beta}$$

とできる。ここで，両辺をｘで微分すれば，

$$\frac{dp}{dx} = e^{c} \beta x^{\beta - 1}$$

が得られ，１日あたりの費用は $e^{c}\beta x^{\beta-1}$ にて近似できる。特に，β～1.0では，e^{c} が１日あたりの費用に相当する。

以上のように，入院費用（＝入院時の病院の収入）は，かなりの部分が入院日数に依存しているが，その依存度は，日数が経つにしたがって漸減していく形を取っていることが判明した。これらの解析結果は，本来，各疾患にわたって詳細に検討すべきであるが，上記のようなデータでも，マクロ的に，病院の収益構造を検討することが可能である。今後，入院医療は，厚生労働省の方針により，包括医療がその主流となるが，包括医療下

においては，このような分析が不可欠なものとなっていくと思われる。

もう一つ重要なことは，予想以上に，医療のばらつきが少なかったということである。つまり，各疾患においての治療のばらつきは，このデータにおいては少ないことが予想できる。

2.3.5 一般化線形モデル

上小節に示したごとく，入院費用は入院日数に著しく依存しているが，他の因子との関連はどうなっているであろうか？ このことを評価するために，入院費用を目的変数とし，説明変数に転帰・治療法・病名大分類・入院日数を含めて，一般化線形モデルによる分析を行った。表3にF値の大きい順に有意な説明変数10個を示した。表に示されたごとく，入院日数によって説明できる部分が大きく，R^2乗値は0.7と高い値を示した。これに対し，二番目以降の属性も入院日数が長期化することに関連すると思われる属性が多く，全体として，手術以外には，入院日数にまさる因子がないことが示された。

表3　対数変換した入院費用に対する一般化線形モデルによる解析結果（上位10項目を示した）

	F-値
入院日数	1590.3
死亡	347.7
精神障害	264.4
妊娠分娩合併症	241.7
循環系疾患	228.2
手術	119.4
周産期発生の病態	98.2
増悪	56.2
治療その他	53.9
神経系疾患	52.1

3．長期時系列データの解析

前節では，病院のデータから病院の特性をマクロ的にとらえるためにデータマイニングの手法を用いたが，病院に蓄積されたデータのもう一つの利用方法はそれを診療支援に役立てることである。特に，10年以上の経過を呈する慢性疾患における病態の変化は今まで検討されてこなかったが，今後，検査データが電子的に蓄積されることで，そのような病態を明らかにする糸口がつかめることが期待できる。

このような目的を掲げて，科研費特別領域研究「データマイニング」(平成13年度～平成16年度)では，上記病院情報システムから抽出された長期時系列データ（通称 肝炎データ）を用いて，11の研究班が，時系列データの解析に取り組んだ[7]。我々のグループでは，時系列医療データの主な特徴は，(1)データの収集間隔と収集期間が個々の患者で異なる不均質系列であること，(2)慢性変化，急性変化など異なる期間で生じるイベントが同一系列中に混在していること，の2点に着目し，これらの課題を克服すべく，(1)構造的類似性に着目した不均質時系列の多重スケール比較法，(2)多重スケール比較とラフクラスタリングに基づくクラスタ分析システムを開発し，慢性肝炎データに適用した。

技術的な詳細については6)を参照していただくとして，本稿ではその成果の概要とその成果を元に慢性疾患の予後を以下に推定したかという追加解析部分について報告する。

3.1 血小板系列の類型化

血小板は，その産生を促進する造血因子（トロンボポエチン）が肝臓で産生されることから，肝

機能の障害状態を反映する指標として注目されている[8]。肝臓の線維化度と血小板数の関係については，各ステージの患者群において血小板数に差がみられ，線維化が進んでいるほど血小板数が減少していることが明らかとなっている[9]。しかしながら，同一患者の血小板数を時系列的に追跡し，線維化の進展と血小板数の減少傾向の関連を調査した例は少なく，また，Ｂ型とＣ型で経過にどのような相違があるかも明らかではない。本研究では，それらを調査する基礎段階として，改良型多重スケールマッチングを用いて血小板数系列の類型化を行い，経時推移にどのような傾向が見られるかを調べた。

3.2 実験手続

対象は，共通データセットに含まれる血小板データ720例のうち，肝生検情報が付随しない222例および検査期間が２週間以下（補間後の構成点数が２点以下）の10例を除外した488例である。類型化の手続きを以下に示す。

１．系列の再構成：各患者の血小板数の時系列を１週間間隔の等間隔時系列に再構成する。最頻検査曜日を基準に１週間ごとサンプリングを行い，当該曜日にデータが存在しない場合は直近検査日のデータにより線形補間する。

２．ウイルス型及びIFN治療の有無によるデータセットの分割：488例のデータをウイルス型に基づきＢ型例とＣ型例に分割し，Ｃ型例をさらにIFN治療の有無に基づき治療有例と治療無例に分割する。各々の例数は，Ｂ型193例，Ｃ型IFN治療有196例，Ｃ型IFN治療無99例である。これらをそれぞれＢ型サブセット，Ｃ型IFN有サブセット，Ｃ型IFN無サブセットと呼ぶ。以降の処理は各サブセットごとに実施する。

３．改良型多重スケールマッチングによる系列比較と相違度行列の作成：サブセットから２つの系列を取り出し，改良型多重スケールマッチングを適用して系列間相違度を算出する。同じ操作をサブセットに含まれる全ての系列組に対して適用し，全系列の系列間相違度をまとめた相違度行列を生成する。

４．階層型クラスタリング法による類型化とクラスタ分析：相違度行列に基づき階層型クラスタリングを適用して樹状図を生成し，クラスタ分析を行う。なお，今回の実験では結合基準として群間平均を用いた。

3.3 実験結果

図８に，Ｂ型，Ｃ型IFN有，Ｃ型IFN無の各

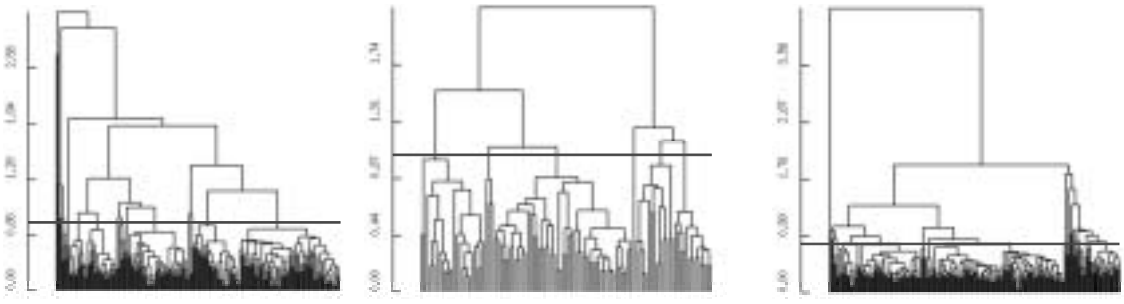

図８　血小板系列の類型化

左：Ｂ型肝炎，中央：Ｃ型（IFN治療なし），右：Ｃ型（IFN治療あり）

サブセットにおいて得られた樹状図を示す。類型の特徴を失わない程度まで併合を進め，それぞれ16，23，6クラスタを得た。同図水平線にて分割位置を示す。この中で，上記分割により生成された各クラスタにおける線維化度別の例構成比（C型IFN有）を表4に示す。表において，最左列はクラスタ番号，続いてF0からF4までの各ステージに層別した例数，最右列はそのクラスタに属する例の総数である。全体的に，線維化度の高い例を多数含むクラスタあるいは低い例を多数含むクラスタのいずれかとなる傾向が強い。クラスタ5，11については，図9，10に具体的な時系列データを示した（クラスタ11は全データの一部）。

以下に得られた知見をまとめる。

1．B型，C型のいずれにおいても，線維化度が高い，特にF4例の構成比が高いクラスタでは，血小板数が慢性的に基準下限を下回る例が多数を占める。また，線維化度がF1あるいはF2でも，F4例と同様に低値推移を呈する例が存在する（B型クラスタ7，C型IFN有クラスタ5）。

2．B型，C型のいずれにおいても，線維化度が低い，特にF1例の構成比が高いクラスタにおいて，血小板数が基準範囲内で推移する例が多数を占める（B型クラスタ5，15，16，C型IFN有クラスタ11，12，23）。また，線維化度がF4でも基準範囲内を維持する例が存在する（同）が，その数は血小板の水準が高いクラスタほど低くなる（B型クラスタ16＞15＞5，C型IFN有クラスタ11＝12＞23）。

3．C型において，血小板数が基準値から持続的に減少を続け，やがて，基準範囲を下回る例がF1，F2を含めて存在する（C型IFN有クラスタ4，6，IFN無クラスタ1）。また，IFN

表4　分類された系列の線維化度別構成比
（C型IFN治療あり）

クラスタ番号	F0	F1	F2	F3	F4	例数
4	0	3	1	0	0	4
5	0	2	1	2	6	11
6	0	1	0	1	1	3
8	0	9	6	9	16	40
9	0	3	0	0	0	3
10	0	1	0	1	3	5
11	2	24	8	9	3	46
12	1	22	10	6	3	42
20	0	4	1	0	0	5
22	0	2	0	0	1	3
23	1	13	4	1	0	19

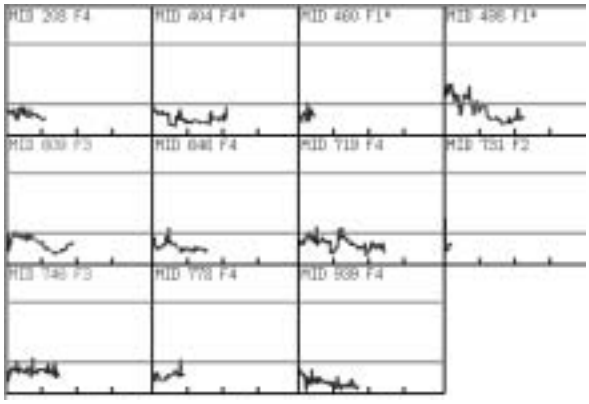

図9　クラスタ番号5の時系列データ
（構成比：F0/F1/F2/F3/F4：0/2/1/2/6）

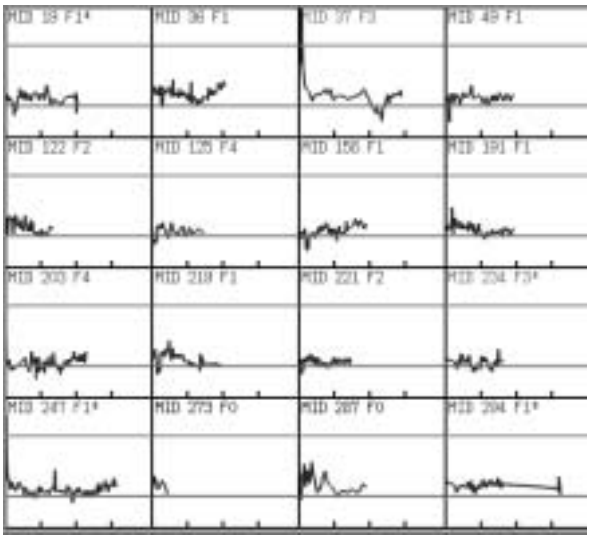

図10　クラスタ番号11の時系列データ（12例）
（構成比：F0/F1/F2/F3/F4：2/24/8/9/3）

治療無の例では，基準下限に至らないものを含めて，より広範に多くの例で減少傾向が見られる（C型IFN無クラスタ１，３)。

４．C型IFN有において，血小板数が基準範囲より低値から持続的に増加を続け，やがて，基準範囲内に至るF４例が存在する（C型IFNクラスタ10)。

これらの結果をもとに，血小板数に基づく肝硬変（F4）までの到達年数とステージ間の経過年数の解析を行った。その結果の一部は３，７）に報告したが，ステージ間の経過年数がほぼ全ての群において平均で１－２年/stage程度であるという結果を得た。他の文献との比較のため年数をベースにした速度に単純に変換すると，例えばC型IFN無の例で1/1.32＝0.76 stage/yearとなり，すでに千葉大のグループから報告されている0.12±0.15 stage/yearと比べるとかなり速く線維化が進むことになる。今回の解析では，(1)血小板数が継続的に基準下限を下回ることをもってF４と見なしていること，(2)増悪が続いて血小板数が増加しない例のみを選択していること，(3)飲酒歴等，個々の事例の背景要素を考慮に入れている訳ではないこと，など，多くの前提を置いているため知見を一般化することはできないが，輸血時などF０を仮定した時点からではなく，ある程度線維化が進んだ状態から同一患者のデータを元にステージ進行に要する経過年数を算出するアプローチにより，興味深い結果が得られたと考える。

４．おわりに

病院情報システムからデータを抽出した後，大学病院の収益に関する特性解析および長期時系列データに関する類型化の方法を報告した。今後，病院情報システムに電子カルテが実装されれば，これらの所見と検査結果との対応付けなど，さまざまな医療に関する問題を検討する機会が到来することになる。データマイニングの手法は，この問題を検討するための極めて有効な技術となりうる可能性がある。また，従来の手法で解決できなかった問題の解決を図る新たな手法を開発するポテンシャルをもった分野であり，今後の発展が期待される。本論文では，近来，実装されるようになった情報系データベース（Data Warehouse）をデータマイニングのフロントエンドとして扱う方法にはふれなかった。Data Warehouseの重要性は，病院情報システムの構築においてもすでに実証されてきている。

島根大学医学部附属病院の病院情報システムにおいても，現在，Data Warehouseを介して，データマイニングを行うシステムを開発しており，今後，この適用例についても報告して行く予定である。

謝辞

本研究は，津本優子が千葉大学医学部大学院在学中の研究結果に拠った。また，本研究後半も，千葉大学医学部附属病院医療情報部からのデータ提供による。千葉大学里村名誉教授および千葉大学医学部附属病院情報企画部高林克日己教授に謝意を表したい。

なお，本研究後半は，文部科学省科学研究費補助金特定領域研究「情報洪水時代におけるアクティブマイニングの実現（領域番号759）の計画研究「ラフ集合に基づくアクティブマイニングによる診療情報生成システムの開発」（課題番号13131208）の助成による。

参 考 文 献

1）里村洋一：電子カルテが医療を変える，日経BP，1998.

2）津本周作：データマイニング技術の臨床応用．最新医学 58(8)：1864－1870，2003.

3）津本周作，平野章二：「複合医工学」としてのデータマイニング．人工知能学会誌 22(2)：201－207，2007.

4）Yuko Tsumoto and Shusaku Tsumoto. Analysis of Hospital Management Data using Generalized Linear Model. Jinglong Wu, Hidenao Fukuyama, Mamoru Mitsuishi, Koji Ito, Shozo Tobimatsu, Toyoaki Nishida (Eds) Complex Medical Engineering, Springer, 2006

5）Shusaku Tsumoto, Shoji Hirano: Automated discovery of chronological patterns in long time-series medical datasets. Int. J. Intell. Syst. 20(7): 737-757 (2005)

6）S. Hirano and S. Tsumoto (2003): Multiscale Analysis of Long Time-series Medical Databases. Proc. AMIA Annual Symposium 2003, Washington DC, 289-293.

7）元田　浩：情報洪水時代におけるアクティブマイニングの実現．平成17年度科学研究費補助金 特定領域研究 研究成果報告書

8）宮崎　洋：トロンボポエチンの将来展望，Jpn J. Transfusion Medicine, Vol.46, No. 3, pp. 311－316 (2000)．

9）Matsumura, H., Moriyama, M., Goto, I., Tanaka, N., Okubo, H., and Arakawa, Y.: Natural course of progression of liver fibrosis in patients with chronic liver disease type C in Japan - a study of 527 patients at one establishment in Japan, J. Viral Hepat., Vol.7, pp. 375-381 (2000)．