データの科学

野　村　義　明

はじめに

現代社会は情報化社会と言われる。実際我々の身の回りにも多くの情報が氾濫している。その大量の情報の中にはいわゆるジャンク情報も大量に存在するのも事実である。一方、信頼できる情報の量も急速なスピードで蓄積されつつあるのも事実である。一年間に発表される科学論文の数も膨大である。MEDLINEの収録論文数は約1,100万といわれている。これら大量の情報を集約し、臨床に役立てていこうというのがEBMの一つのコンセプトであると思われる。EBMは基礎研究のみでなく、臨床研究こそ重要な研究であることを示してくれた。多くの臨床医にとって各自の所有する患者データを一定のフォーマットにまとめることによって日常の臨床も充分にサイエンスとして価値のあるものであることを示してくれたという点でその功績は多大なものである。しかし、EBMはあくまでも集団に対する平均値を示したものであり、データを目の前の患者に適用する際には個別化と称し医療のアートの部分に臨床医の裁量権に自由度を残している。この点はEBMでも議論のあるところで、単に論文データを目の前の患者に適用する際に集団の平均値をあるサブグループに適用する危険性があると指摘されている。

【著者連絡先】
〒230-8501 神奈川県横浜市鶴見区鶴見2-1-3
鶴見大学歯学部　予防歯科学講座　野村義明
TEL：045-581-1001　FAX：045-573-9599

データマイニング

一方、前述のように情報化社会においては大量の情報を集積し、解析することによって企業戦略等に役立てている。特にインターネットによる情報収集には著しい進歩がみられる。インターネットによる購買やアンケート調査においては、個人情報も大量に収集されている。また、ある企業では、マウスクリックの軌跡まで解析することによって各個人の趣向性を解析し、商品のレコメンデーションやホームページデザインに役立てている。この方法はWebマイニングを呼ばれデータマイニングの一手法として位置づけられている。このように大量のデータからある一定のルールを見つけ出し企業の販売戦略や企業経営に役立てる手法をデータマイニングという。

データマイニングの利用例としてはバスケット分析による商品陳列や、ウエッブ上でのレコメンデーションが挙げられる。バスケット分析とは客がどのような商品を組みあわて購買するかを分析する方法で、どの商品を同じ買い物かご（バスケット）に入れるかを分析する手法である。現在は購買時の会計がバーコードで行われることが多いためデータが大量に存在する。基本的には単純なクロス集計であるが、商品の購買を分析する場合検討する項目数も膨大である。例えば100品目あれば5050回、1000品目あれば500500回のクロス集計を行うことになる。この解析を一週間単位で行うとすれば、一人二人の人間が行う業務内容を遙かに逸脱している。このような解析を一気に行ってしまおうというのがバスケット分析であ

る。現在はYes，No型の2値データしか利用できないこと、また分析自体は頻度による分析のみでχ^2乗検定や正確確率を求めているわけではないので統計学的な厳密性には欠けるといった欠点は存在するが大量のデータを一気に処理するといった点で利用価値は高い。結果はWebグラフ、蜘蛛の巣グラフといった形で表現される。この解析方法を利用して歯牙の欠損パターンの解析を行った実例を図1に示す。クロス集計を528回繰り返さなければならない手間を一気に解決してくれる。

齲蝕発症予測モデルの構築

データマイニングの特徴としては、大量のデータを扱うこと、利用できる解析手法を複数用い最も適したモデルを選択するといったことであろう。

この点において、医科領域では多変量解析の手法としてロジスッテク回帰分析がビジネスにおける解析においてもロジステック回帰分析は主要な解析方法の一つであるが、この手法に限ることなくDecision Analysisやニューラルネットといった解析方法を常用し、これら複数のモデルから最も適したモデルを選択するといたことを行っている。

齲蝕の病因は多因子性であり、その解釈には疫学の三角形のモデルが適用され、カイスの3つの輪として知られている。歯科領域は、対象とする疾患が齲蝕、歯周病がそのほとんどをしめ、これらの疾患は罹患率の高い疾患である。またその処置も通常は一次医療機関で可能であり、医療機関の数も開業医が多く2次、3次医療機関が極端に少ないといった特徴がある。齲蝕罹患率の減少、かかりつけ歯科医院の推奨等から歯科医療全体が従来の齲蝕治療を中心とした治療から予防を中心とした患者管理へ移行しつつある。一次医療機関である開業医においては診療所に所属する歯科医師が固定されており一人の患者の長期観察が可能であり、また長期間にわたるデータが蓄積されて

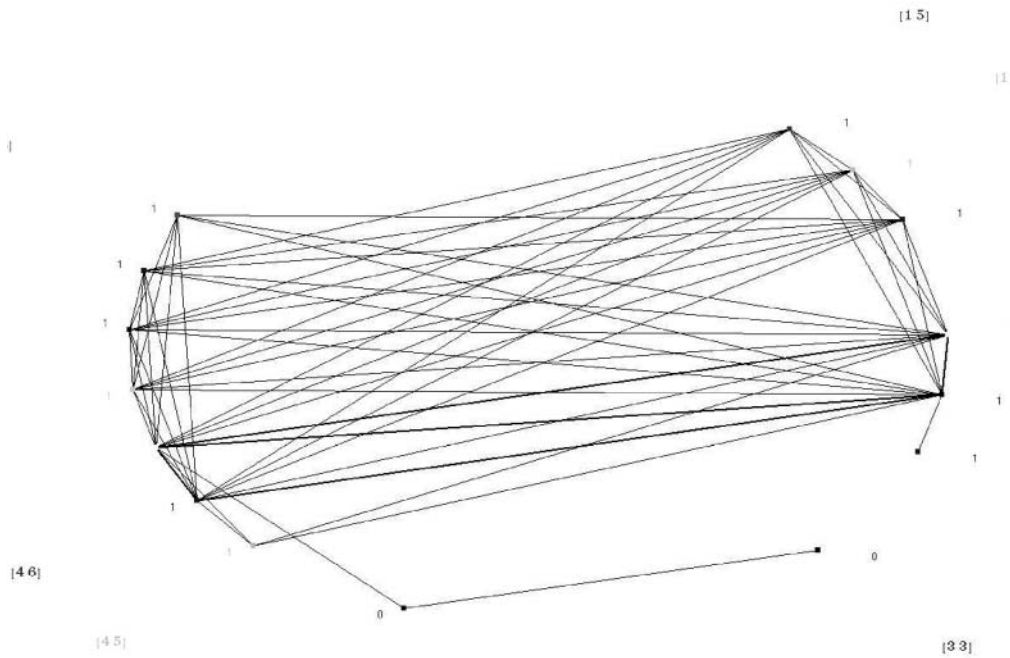

図1　高齢者344名の欠損パターンの解析
点線の部分は344症例中200例以上のみられた欠損の組み合わせ
太線の部分は344例中240症例以上にみられた欠損の組み合わせ

いるケースが多く存在する。

齲蝕に関しては「リスク検査」と称した検査システムが確立されており、データに基づく患者管理を実践している開業医も多く存在する、またこれらの医療機関には大量の患者データが蓄積されている場合が多い。歯科診療所通院患者のうち、年齢が20歳以下で齲蝕治療を終了し口腔衛生指導を行った患者2131名のデータの中から、初診時、治療終了時、メインテナンス時にリスク因子の評価を行った1664名の患者データを使用し、メインテナンス時に新規に齲蝕が発症した患者と発症しなかった患者を比較検討した。従来から齲蝕のリスク因子とされてきたミュータンスレンサ球菌量、乳酸桿菌の量、5分間唾液流出量を測定し、唾液緩衝能、フッ素の使用状況、食事回数および乳歯齲蝕の経験歯数を検討項目とした。解析方法は新規齲蝕発症した患者と発症しなかった患者に分け、データのなかからその50％を無作為に抽出し、ロジステック回帰分析、ニューラルネットワーク、決定木分析による齲蝕発症モデルを作成を複数作成した。作成したモデルに対しさらにデータの50％無作為に抽出しモデルの妥当性を検討し、最も予測精度の高いモデルを各解析方法から選び、3種のモデルの予測精度を検討した。ロジステック回帰分析による各項目のオッズ比を表1に示す。ロジステック回帰分析ではHosmerとLemeshowの検定で有意確率が0.05以下でモデルの適合が得られなかった。決定木分析で最も予測精度が高かったモデルを図2に示す。ニューラルネットワークでは入力層、隠れ層、出力層それぞれ9，4，3ニューロンのモデルが構築され各因子の相対重要度を表2に示した。

表1

	オッズ比	95%信頼区間	P-value
乳酸桿菌量	1.002	0.78-1.29	0.989
唾液緩衝能	1.508	1.08-2.11	0.016
5分刺激唾液量	1.129	1.01-1.27	0.037
乳歯齲蝕の経験歯数	1.676	1.23-2.29	0.001

また各モデルの予想精度を繰り返し検討した結果、ロジステック回帰分析で52.2％、決定木分析

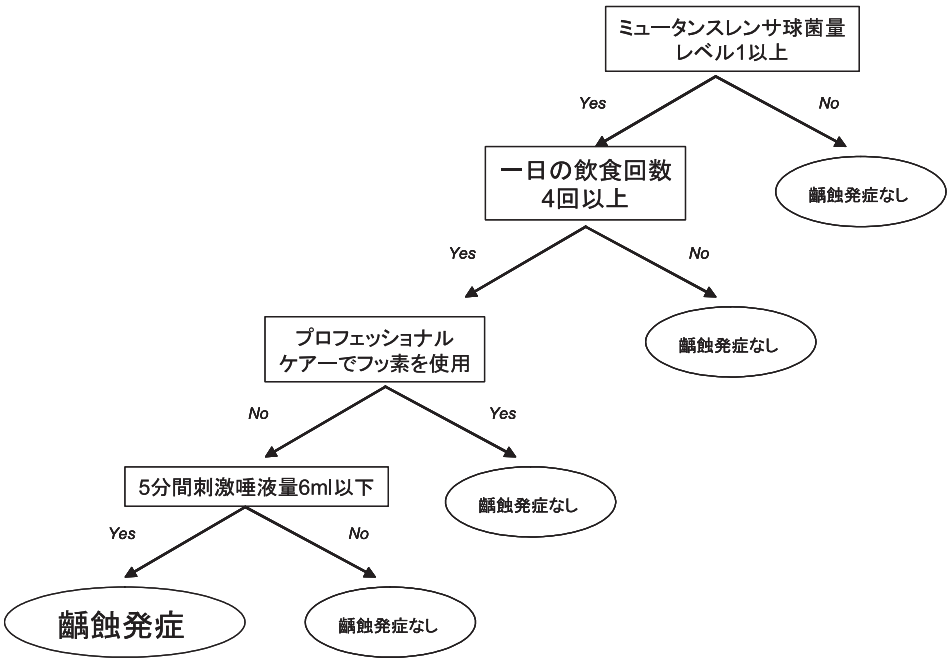

図2

表2

	相対重要度
5分間刺激唾液量	0.49
乳歯齲蝕の経験歯数	0.36
乳酸桿菌量	0.24
ミュータンスレンサ球菌量	0.17
プラーク量	0.15
一日の飲食回数	0.07
プロフェッショナルケアーでのフッ素の使用	0.05
ホームケアーでのフッ素の使用	0.02
唾液緩衝能	0.02

で69.5％、ニューラルネットワークで71.6％であり、ロジステック回帰分析による予想精度が最も低かった。決定木分析やニューラルネットワークといった手法の導入は今後の解析においても有用である。

現時点ではこれら3つの因子のうち宿主の因子に対する評価方法があまりにも未熟である。唯一客観的な評価方法が唾液流量や唾液緩衝能の評価でありその他の評価方法といえばフッ素の使用状況の問診等であることからもその事実は充分に伺える。現在はバイオインフォマテックスの進歩により、ヒトゲノムのシークエンスは明らかになりポストゲノムの時代とも呼ばれている。未だにシークエンスの明らかとなった遺伝子の約半分はその機能が明らかにはなっていないという現実があるものの［1]、プロテオミックス等の技術進歩により近い将来その機能が解明されることが期待される［2]。このような技術をもって将来齲蝕発症や歯周病発症に関連する遺伝子、タンパク質が解明されれば宿主の因子に対する評価が大きく変化するものと期待している。

まとめ

以上、データマイニングの手法を用いた解析の実例を示したが、平均値を提示するEBMで厳しく制限されていたサブグループ解析をデータマイニングではセグメンテーションと称し日常的に行っている。企業経営としても商品を購買しない顧客にダイレクトメール等を出すことは経費の無駄である。この観点から医療においても大量のデータがあれば治療効果の高い患者を選別することが可能になる。前述のように歯科では対象とする疾患が限られていること、これらの疾患が罹患率が高い疾患であることを考慮するとデータマイニング手法の応用には限りない可能性を秘めているといえよう。現在、データ収集のためにウエッブ上でのデータの大量収集システム等を構築してゆくことを計画中である。

文　献

1）磯辺俊明　プロテオミックスの動向と企業戦略　バイオベンチャー　Vol1　No3　32-37　2001

2）MacBeath G, Schreiber SL. Printing proteins as microarrays for high-throughput function determination Science. 289：1760-3　2000