様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年　5 月 28 日現在

機関番号：１６４０１
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：２１５９０５６６
研究課題名（和文）　病院情報システムのデータ活用方法の確立とメディカルデータマイニングの創成

研究課題名（英文）　Establishment of the practical data utilization method of hospital information systems, and creation of medical data mining

研究代表者
奥原　義保（OKUHARA YOSHIYASU）
高知大学・教育研究部医療学系・教授
研究者番号：40233473

研究成果の概要（和文）：病院情報システムには日々の診療に伴って膨大なデータが蓄積されているが、直接の診療以外にはほとんど有効に２次利用されていない。この潜在的に極めて有用と考えられるデータを、医学研究に活用するための方法を、具体的課題の解析によって示した。特に、長期にわたる検査値データの変動補正、検査診断特性の網羅的解析、血球検査データの検査診断特性パターンによる疾患診断、検査値時系列の計算知能による特徴抽出と予測、などについて成果を得た。

研究成果の概要（英文）：Nowadays, a vast amount of data is stored in hospital information systems with daily medical practices. However, utilization of such data is limited to the direct medical care for patients themselves. The indirect usage for medical studies is still not so active. We have shown various examples of the effective data analysis using accumulated medical data in a hospital information system. In particular, Effective calibration method for laboratory test data for long time period , Comprehensive analysis of the diagnostic characteristic pattern for laboratory test data, Diagnostic characteristics for CBC cell distribution data, Prediction by pattern extraction with artificial intelligence from laboratory-test time series data are shown to be useful.

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	800,000	240,000	1,040,000
2011 年度	700,000	210,000	910,000
年度			
年度			
総　計	2,900,000	870,000	3,770,000

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、データサイエンス、データマイニング、病態推移予測、予防医学

１．研究開始当初の背景

近年の我が国における病院情報システムの普及に伴い、膨大な医療データが多くの医療施設に蓄積されつつある。しかしながら、

病院情報システムのデータ利用は、ほとんどが個々の診療における直接の利用、すなわち１次利用に限られており、データの蓄積を２次利用するという面での有効活用はほとんど行われていない。

こうした状況は国内に限らず、国外においても同様であり、薬剤疫学の分野で一定の成果が見られるものの、それ以外の応用については限定的かつ散発的で、解析の方法論は確立していない。

一方で、こうしたデータの蓄積を、様々な医学的研究に有効に活用することが期待されている。例えば、大量のデータに基づく「予測」、「評価」、「診断支援」、「稀な事象の検出」などへの活用が考えられる。特に、長い期間にわたるデータの蓄積がある場合は、長期的な観察が必要な生活習慣病を対象にした、疾病の発症予測や危険因子の探索、新たな機序の推測などの研究や、薬効の長期的評価、薬剤の有害事象の検出などの研究への利用が考えられる。

また、こうしたデータは既に電子化されて蓄積されているため、原理的には、柔軟なアイデアに基づくデータの自由な組み合わせにより、様々な切り口から探索的な解析を迅速に行なうことが可能であり、有効な方法さえ確立すれば、疫学をはじめとする集団基盤的な医学研究において、従来は不可能であった速度で多岐に渉るテーマを推進し、研究の進め方と生産性において画期的な変革をもたらすことが予想される。その結果、予防医学や医療政策へのタイムリーな提言においても多大な貢献ができると考えられる。

このような大量の電子データを解析するための実際の手法としても、コンピュータの急速な性能向上に伴い、数理統計解析の手法やデータマイニングの各種手法などの強力な方法が揃いつつある。

しかしながら、病院情報システムのデータ解析には、一般的なこれらの手法がそのまま使えるわけではない。日々の診療の結果蓄積されたデータであるため、複雑な多様性や情報の不完全さ、データ量の偏在などを含む。このため、解析においては、事前の計画に基づいて収集されたデータの場合とは異なる手法が必要である。

したがって、個々の事例に応じてデータ特性を見極め、使えるデータの選択と質の評価、欠損値の処理等の事前作業、いわゆるデータクレンジングが必要である。こうした作業の後、データ特性に応じた適切な方法を適用する必要がある。さらに、有効性が確認された手法を体系化しシステマティックな方法論を確立してゆくことが必要である。

すなわち、医学を軸に、数理統計学、情報科学、情報工学等を基にした分野横断的なアプローチにより、「メディカルデータマイニング」とでも言うべき、医療データに特化した新しい解析の体系を開拓することが急務である。

こうした状況の中で、高知大学の総合医療情報システム IMIS(Integrated Medical Information System)には２７年という長期間に渉って２４万人以上の臨床データが蓄積されており、単独の施設としては我が国で最も規模の大きい医療データの蓄積があるなど、質・量両面において、こうした解析のために有利な条件を備えている。

２．研究の目的

この研究では、IMIS に蓄積されている医療データを対象に、統計学的手法やニューラルネット、その他のデータマイニング手法等の様々な解析方法を用い、大量の医療データに基づく「予測」、「評価」、「診断支援」、「稀な事象の検出」を目的としたいくつかの具体的課題について、有用な結果を導出し、有効な方法の確立を目指す。

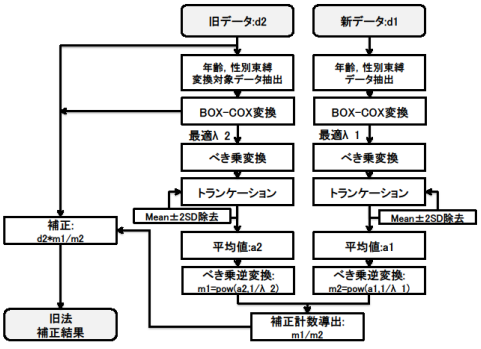
３．研究の方法

本研究では、様々なテーマにつき、様々な方法を試みたが、以下、特に良好な結果が得られた方法につき説明する。

（１）長期間検査値データの変動補正

長期間にわたる検査データの蓄積を２次利用する場合は、検査方法変更に伴う検査値変動の影響を補正することが、解析の信頼性を保証するために必須となる。本研究では、すべての検査値データ解析の基本となるデータクレンジング手法の一つとして、検査方法変更前後の検査値の分布をそれぞれ BOX-COX 変換によって正規分布に直した上で、平均値が変更前後で連続になるよう接続し、逆変換によって元に戻すという方法の有効性を確認した。

検査方法変更前後のデータ接続



（２）検査診断特性の網羅的解析

病名歴と検査歴に対して ROC 解析の手法を用いて網羅的解析を行い、その結果を知識データベース化することにより、ある病名の診

断に適した検査の検索や、ある検査で診断可能な病名、ある患者の検査値のセットから考えられる病名をリストアップすることなどを可能にすることを試みた。

（３）CBC 検査データの検査診断特性パターンによる疾患診断

サイトフローメトリーにより血液の検査(CBC:Complete Blood Count)を行うと、各血球に照射されたレーザー光の前方散乱強度、側方散乱強度、蛍光散乱強度から、それぞれ、細胞の大きさ、細胞の内部構造、細胞内の DNA,RNA 量などを知ることができる。これらの強度の組み合わせを 2 次元の散布図にプロットし、そのパターンを解析することにより、多くの情報を得ることができる。

2006 年 4 月より，サイトフローメトリーによって計測される血球細胞情報を収集した，約 50 万件のデータを用い、一人ずつの情報を 32×32 の 2 次元散布図 (1024 項目の検査) に変換・凝縮し、病名歴と連動させて網羅的に ROC 分析を実施、良好な診断特性を示すパターンを探索した。

（４）検査値時系列の計算知能による特徴抽出と予測

血清 ChE 値が肝予備能を反映する良い指標であることを利用し、IMIS に蓄積された肝硬変患者の ChE データを対象として、発症初期の実測値からどの程度その後の ChE 値の推移が予測できるかの研究をおこなった。解析用データウェアハウスに記録されているデータは、日常診療に伴うデータであり、必ずしも一定間隔の定期的な検査データではないため、そのまま解析するには困難が伴う。その問題を解決するため、Artificial Neural Network によるモデル化を行い、一定間隔の検査データを生成し、欠損値を補間することで解決した。

また、各患者データにおいては、個人差のため、基準値がそれぞれ異なっており、直接比較することが困難である。このため、モデル化データの変動量に注目して、各患者ごとの比較を行った。

肝硬変患者において ChE 値が低下している期間の変化量パターンの分類を行い、そのパターン及び各患者における ChE 値に基づき、低下量を予測するアルゴリズムを構築した。

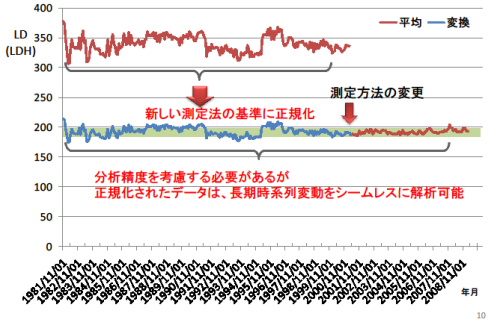
４．研究成果

（１）長期間検査値データの変動補正

検査方法変更前後の検査値の分布をそれぞれ BOX-COX 変換によって正規分布に直した上で、平均値が変更前後で連続になるよう接続し、逆変換によって元に戻すという方法は、高知大学の 30 年間にわたる検査値の蓄積においても、精度を保証できる有効な方法であることが示された。

この方法により、長期間にわたる検査値の蓄積を精度を保証しつつ様々な解析に用いることができるようになり、疫学をはじめとする集団基盤的な医学研究において、有用なデータを提供することが可能になると考えられる。

正規化処理結果



（２）検査診断特性の網羅的解析

検査歴データと病名歴データの関係を網羅的に解析し、初期のスクリーニング診断に適した検査診断能のデータベース（知識データベース）を構築した。この知識データベースを利用することにより、ある病名からその病名の診断に有用な検査項目を検索することができる。また、ある検査から検査診断特性の高い病名を検索できる。初回の確定診断日から 14 日前までの直近検査データから導出した知識データベースでは、299 病名、751 項目の検査に対する診断特性が生成できた。さらに、30 日前までの検査に対しては、4206 病名、1350 項目の検査診断特性が生成できた。これらの知識ベースを探索することで、これまで知られていなかった検査の診断特性を発見でき、診断支援システムの構築も可能である。また、医学教育においても、有用な教育用システムの構築に応用できる。

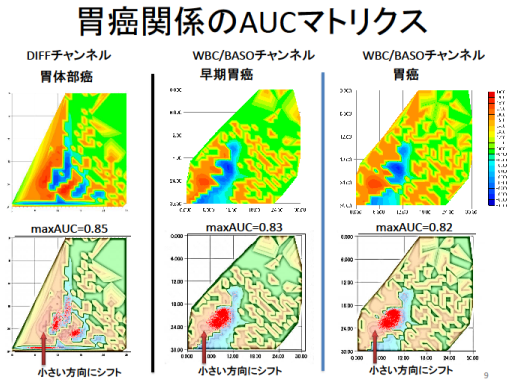
陽性尤度比による情報提示

項目名称	検査値	糖尿病性網膜症	腎不全	糖尿病
GLU	195		1.335	17.544
T-CHO	328		∞	∞
CHE	329	2.139	0.587	1.352
CRN	2.91		∞	10.492
BUN	30	1.948	20.599	3.001
UA	5.4	1.156	2.180	1.285
AMY	122		2.616	1.178
S-K	5	1.569	5.527	1.214
S-CL	111	2.750	6.621	2.279
SEGMENT	74	0.163	2.199	1.252
尿糖	500	5.797	0.855	76.296
尿蛋白	1000		∞	∞
尿上皮	25		4.480	2.993
尿円柱	5		9.935	2.138
U-R/HPF	34		3.310	14.502
U-W/HPF	8		3.417	2.910
Hb-A1c	6.8	2.071	0.958	21.964

（３）【CBC 検査データの検査診断特性パターンによる疾患診断】

これまで蓄積された CBC 検査データを網羅的に精査することにより、血液学分野以外にも、数多くの良好な検査診断特性を示すパターンが存在することを明らかにした。例えば、胃癌の診断が ROC 分析で感度=74%，特異度=88%，AUC=0.85 と，良好な検査診断特性を示す結果が示された。また、多次元的なパターンによる解析手法を確立し、早期胃癌を含む胃癌のパターンは小児や妊娠期に類似したパターンを示すが、独立したパターンであることを発見した。

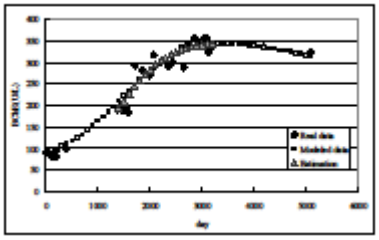
この方法を、癌をはじめとした様々な疾患の早期発見が可能な臨床検査方法として確立したい。特に、感染症パンデミックの早期検出方法として有望であると考えられる。



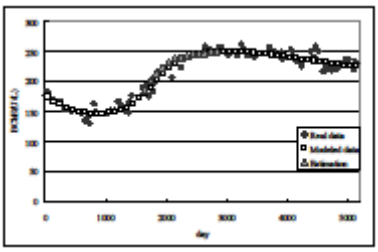
（４）検査値時系列の計算知能による特徴抽出と予測

118 例の、インターフェロン治療が行われていない肝硬変患者のデータを学習データとし、別の 32 例の肝硬変患者（インターフェロン治療を行っている症例も含む）を用いて方法の評価を行った。代表的なパターンを以下に示す。

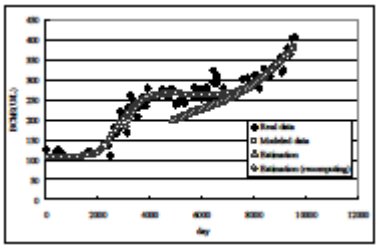
①インターフェロン治療が行われていない肝硬変患者の予想（最後の検査より遡って、5100 日前から 3100 日前の実測データによる）は、良く一致する。



②ChE値が既に低くなった状態でのインターフェロン治療開始例の予測（5100日前から2900日前の実測値による）も、良く一致する（インターフェロンの効果が見られない）。



③ChE値がまだ高い状態でのインターフェロン治療開始例（7412日前から開始）は、7300日前から4800日前の実測データによる予測とは一致しない（インターフェロンの効果が見られ、予測より良い方向に回復する）。インターフェロン投与後の実測値（4700日前から5700日前）を加えて再計算すると、良く一致する。



全体として、実際の検査データ及び予測データとの平均誤差は、③のような再計算無しで18.4であった。このことから、肝硬変患者のChE値の時間的振る舞いは、本研究で提案した方法によれば、比較的初期のデータを用いて予測できることが分かった。

この方法による同様の予測は、一般的に長い時間をかけてゆっくり変化する検査値に対して有効であると考えられ、慢性疾患の病態推移予測に応用できると考えられる。特に、日常診療に伴うデータのように、検査の間隔が一定ではなく、欠損の多いデータへの応用に適している。

本研究においては、ここに説明した例以外にも、「画像情報を用いた病態分類」、「キャピラリー電気泳動分析から得られる波形データの解析」、「糖-インスリン動態の数理モデルを用いた糖尿病発症予防」、「状態空間モデルによる病態推移予測」、「NSAIDsのCOX選択の有無による急性腎機能障害のリスクに関する研究」などのテーマについても解析を進めており、今後も、成果の発表を続け、病院情報システムのデータ活用方法をより確かなものにするとともに、医療データからの発見を行うメディカルデータマイニングを発展させてゆくつもりである。

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）

〔雑誌論文〕（計３件）

① Yutaka Hatakeyama, Hiromi Kataoka, Noriaki Nakajima, Teruaki Watabe, and Yoshiyasu Okuhara,” Algorithm for Estimation of Thyroid Gland Size in Ultrasonography Images for Extracting Abnormal Thyroid in Medical Practice,” Journal of Advanced Computational Intelligence and Intelligent Informatics, 査読有、Vol.16 No.1,2012, pp87-93,

② Yutaka Hatakeyama, Hiromi Kataoka, Noriaki Nakajima, Teruaki Watabe, Yoshiyasu Okuhara. ″A Classification Algorithm of Abdominal Ultrasound Images in Medical Practice for Secondary Uses ″. J. of Advanced Computational Intelligence and Intelligent Informatics. 査読有,14,2010,128-134

③ Yutaka Hatakeyama, Hiromi Kataoka, Noriaki Nakajima, Teruaki Watabe, Yoshiyasu Okuhara, ″Estimation Algorithm of Butyrylcholinesterase for Cirrhosis using Neural Network″. IC-MED. 査読有、3,2009,77-86

〔学会発表〕（計１８件）

① Hiromi Kataoka, : CBC cell distribution data yield remarkable laboratory diagnostic characteristics, The 8th International Conference of Clinical Laboratory Automation, (Kora,Seoul), April 12, 2012

② 片岡浩巳、診療データベースの全組み合わせから計算した 診断能データベースシステムの構築,日本医療情報学会,第31 回医療情報学連合大会,鹿児島,2011/11/22

③ 片岡浩巳,大規模データを用いた知見の発見,日本医療情報学会,第 31 回医療情報学連合大会,鹿児島,2011/11/21

④ 渡部輝明、”糖化ヘモグロビン検査値における生理的変動の確率論的振る舞いとその評価方法”,第31回医療情報学連合大会、鹿児島11月21-23日、2011年

⑤ 畠山 豊,” 電子カルテ上の手術履歴を考慮した退院サマリ分類アルゴリズム,” 日本知能情報ファジィ学会、ファジィシステムシンポジウム 2011、福井、2011/9/12-14

⑥ 奥原義保、今後の医療データ解析に求められる医療情報学の役割と人材育成、日本医療情報学会課題研究会シンポジウム「患者の視点に立った医療データ分析に関する研究」、2011年11月22日，東京（コクヨホール）

⑦ 畠山豊、病院情報システム上における血液ガス評価システム，日本医療情報学会、第30回医療情報学会連合大会,2010年11月20日.アクトシティ浜松（浜松）

⑧ Hiromi Kataoka ,Quantitative electrophoretic profiling for diagnostic prediction of major serum constituents,AACC,2010 年 7 月 25 日,Anaheim Convention Center、アナハイム（米国）

⑨ Yutaka Hatakeyama, Acid-base Balance Evaluation System based on Fuzzy Inference, World Automation Congress (WAC), International Forum on Multimedia and Image Processing (IFMIP) 2010, 2010年9月12日，神戸国際会議場（神戸）

６．研究組織

(1)研究代表者
奥原　義保（OKUHARA YOSHIYASU）
高知大学・教育研究部医療学系・教授
研究者番号：40233473

(2)研究分担者
畠山　豊（HATAKEYAMA YUTAKA）
高知大学・教育研究部医療学系・准教授
研究者番号：00376956
渡部　輝明（WATABE TERUAKI）
高知大学・教育研究部医療学系・講師
研究者番号：90325415
中島　典昭（NAKAJIMA NORIAKI）
高知大学・教育研究部医療学系・助教
研究者番号：00335928
片岡　浩巳（KATAOKA HIROMI）
高知大学・教育研究部医療学系・助教
研究者番号：80398049