データマイニング手法を用いた糖尿病臨床情報の解析

16615006

平成16年度～平成17年度科学研究費補助金

（基盤研究（C））研究成果報告書

平成19年3月

研究代表者　佐倉　宏

東京女子医科大学医学部講師

平成 16～17 年度科学研究費補助金（基盤研究（C））により、「データマイニング手法を用いた糖尿病臨床情報の解析」の研究を行い、ここにその成果を報告する。なお、東京女子医科大学大学院生　春木武徳氏には本研究に多大に協力していただいた。また、検査技師の田中康富氏にはデータ入力・解析に協力していただいた。

研 究 組 織

研究代表者：佐倉　宏（東京女子医科大学医学部講師）
研究分担者：菅野宙子（東京女子医科大学医学部助手）
研究分担者：丸山聡子（東京女子医科大学医学部助手）
研究分担者：宇都祐子（東京女子医科大学医学部助手）
研究分担者：岩本安彦（東京女子医科大学医学部教授）
（研究協力者：春木武徳（東京女子医科大学大学院生））

交付決定額（配分額）　　（金額単位：円）

	直接経費	間接経費	合計
平成 16 年度	2,100,000	0	2,100,000
平成 17 年度	1,600,000	0	1,600,000
総　計	3,700,000	0	3,700,000

研 究 発 表

(1) 学会誌等

1. 佐倉　宏、岩本安彦、OLAP/データマイニングを用いた糖尿病入院患者の血液改善に関する因子の解析、糖尿病 47(Suppl. 1) S-72, 2004 年 5 月
2. 菅野宙子、佐倉　宏、丸山聡子、宇都祐子、藤川径子、岩本安彦、データベースを用いた経口血糖降下剤の効果の解析、糖尿病 47(Suppl. 1) S-162, 2004 年 5 月
3. 丸山聡子、佐倉　宏、菅野宙子、宇都祐子、藤川径子、田中康富、岩本安彦、血糖コントロールを主目的に入院した患者の HbA1c 改善度に関与する因子の解析、糖尿病 47(Suppl. 1) S-297, 2004 年 5 月
4. 藤川径子、佐倉　宏、菅野宙子、丸山聡子、宇都祐子、岩本安彦、食事負荷試験を用いた血糖とインスリン分泌の解析、糖尿病 47(Suppl. 1) S-242, 2004 年 5 月
5. 佐倉　宏、春木武徳、岩本安彦、電子カルテから抽出した糖尿病初診患者情報を用いた血糖コントロール改善因子の解析、第 25 回医療情報学連合大会（第 6 回日本医療情報学会秋期学術大会）プログラム・抄録集、130、2005 年 11 月

6. 春木武徳、佐倉　宏、岩本安彦、電子カルテテンプレートを用いた初診情報のデータマート化、第 25 回医療情報学連合大会（第 6 回日本医療情報学会秋期学術大会）プログラム・抄録集、129、2005 年 11 月
7. 佐倉　宏、菅野宙子、岩本安彦、経口薬治療で血糖コントロールはどの程度達成されているか？（２型糖尿病の新しい薬物療法)、糖尿病 48(Suppl. 2) S-45, 2005 年 5 月
8. 菅野宙子、佐倉　宏、藤川径子、田中康富、岩本安彦、スルホニル尿素薬およびナテグリニドの薬物効果の解析、糖尿病 48(Suppl. 2) S-54, 2005 年 5 月
9. 春木武徳、佐倉　宏、菅野宙子、丸山聡子、藤川径子、田中香野、岩本安彦、多変量解析を用いた初診糖尿病患者の HbA1c 改善寄与因子の検討、糖尿病 48(Suppl. 2) S-169, 2005 年 5 月
10. 田中香野、佐倉　宏、丸山聡子、菅野宙子、宇都祐子、藤川径子、春木武徳、田中康富、岩本安彦、インスリン導入後の血糖コントロールを予測し得る、患者の生理的・心理的因子、糖尿病 48(Suppl. 2) S-59, 2005 年 5 月
11. 大倉　毅、佐倉　宏、佐藤麻子、中神朋子、谷口晋一、重松千秋、岩本安彦、インスリングラルギン外来導入 2 型糖尿病症例の検討、糖尿病 48(Suppl. 2) S-135, 2005 年 5 月

(2) 口頭発表

1. 菅野宙子、佐倉　宏、丸山聡子、宇都祐子、藤川径子、岩本安彦、データベースを用いた経口血糖降下剤の効果の解析、第 47 回日本糖尿病学会年次学術集会、2004 年 5 月 14 日
2. 佐倉　宏、岩本安彦、OLAP/データマイニングを用いた糖尿病入院患者の血液改善に関する因子の解析、第 47 回日本糖尿病学会年次学術集会、2004 年 5 月 15 日
3. 藤川径子、佐倉　宏、菅野宙子、丸山聡子、宇都祐子、岩本安彦、食事負荷試験を用いた血糖とインスリン分泌の解析、第 47 回日本糖尿病学会年次学術集会、2004 年 5 月 15 日
4. 丸山聡子、佐倉　宏、菅野宙子、宇都祐子、藤川径子、田中康富、岩本安彦、血糖コントロールを主目的に入院した患者の HbA1c 改善度に関与する因子の解析、第 47 回日本糖尿病学会年次学術集会、2004 年 5 月 15 日
5. 田中香野、佐倉　宏、丸山聡子、菅野宙子、宇都祐子、藤川径子、春木武徳、田中康富、岩本安彦、インスリン導入後の血糖コントロールを予測し得る、患者の生理的・心理的因子、第 48 回日本糖尿病学会年次学術集会、2005 年 5 月 12 日
6. 菅野宙子、佐倉　宏、藤川径子、田中康富、岩本安彦、スルホニル尿素薬およびナテグリニドの薬物効果の解析、第 48 回日本糖尿病学会年次学術集会、2005 年 5 月 12 日
7. 春木武徳、佐倉　宏、菅野宙子、丸山聡子、藤川径子、田中香野、岩本安彦、多変量解析を用いた初診糖尿病患者の HbA1c 改善寄与因子の検討、第 48 回日本糖尿病学会年次学術集会、2005 年 5 月 13 日

8. 大倉　毅、佐倉　宏、佐藤麻子、中神朋子、谷口晋一、重松千秋、岩本安彦、インスリングラルギン外来導入2型糖尿病症例の検討、第48回日本糖尿病学会年次学術集会、2005年5月13日
9. 佐倉　宏、菅野宙子、岩本安彦、経口薬治療で血糖コントロールはどの程度達成されているか？（2型糖尿病の新しい薬物療法)、第48回日本糖尿病学会年次学術集会、2005年5月14日
10. 佐倉　宏、春木武徳、岩本安彦、電子カルテおよび情報システムを利用した糖尿病臨床データ解析、第5回糖尿病教育資源共有機構年次学術集会、2005年8月6日
11. 春木武徳、佐倉　宏、岩本安彦、電子カルテから抽出した糖尿病初診患者情報の解析、第5回糖尿病教育資源共有機構年次学術集会、2005年8月6日
12. 佐倉　宏、大規模データベースから学ぶ経口糖尿病薬の使い方 ~経口薬治療で血糖コントロールはどこまで達成できるか、第４回経口糖尿病薬フォーラム、2005年11月19日
13. 春木武徳、佐倉　宏、岩本安彦、電子カルテテンプレートを用いた初診情報のデータマート化第25回医療情報学連合大会／第6回日本医療情報学会秋期学術大会、2005年11月25日
14. 佐倉　宏、春木武徳、岩本安彦、電子カルテから抽出した糖尿病初診患者情報を用いた血糖コントロール改善因子の解析、第25回医療情報学連合大会／第6回日本医療情報学会秋期学術大会、2005年11月25日
15. 佐倉　宏、糖尿病におけるITの活用、第28回糖尿病治療研究会、2006年2月25日

(3) 出版物

1. 佐倉　宏、糖尿病治療の新たな展望、糖尿病とデータマイニング、からだのの科学（増刊）糖尿病2005, 287-292, 2004年7月25日

研究成果による工業所有権の出願・取得状況

なし

研究目的

①分子生物学が発達し、糖尿病の分野においてもMODY（若年発症2型糖尿病）をはじめとする特殊なケースの原因遺伝子が解明された。申請者自身もインスリン、インスリン受容体、ミトコンドリア、グルコキナーゼ遺伝子の遺伝子異常を同定してきた。しかし、なお95％以上の糖尿病の原因遺伝子は未知である。Association analysisやsib-pair analysisなどでは原因遺伝子が見出せないでおり、依然として「糖尿病は遺伝学の悪夢」の状態が続いている。この理由として、糖尿病の発症・進展・治療には環境因子が多大な影響を及ぼしているが、遺伝子解析には環境因子があまり考慮されていないことが主因であると考えられる。また、糖尿病の分野ではリサーチエビデンスに乏しく、EBMの実践が困難な領域であるが、これも環境要因の扱いが困難なためと考えられる。

そこで、本研究においては、①環境要因を含めた糖尿病臨床情報をデータベース化すること、②従来の疫学的解析手法に加えて、データマイニング手法を用いて、糖尿病の発症・進展・治療に関与する因子を分析することを目的とする。具体的には、経口血糖降下薬の効果の解析及び初診および入院患者の予後解析を行う。特に環境要因としては、食事摂取状況、運動状況、患者の心理的状況、医師・看護師による患者評価、さらに、細小血管障害、大血管障害、日常生活機能、社会生活機能を総合したQOLについて、患者評価の指標に加える。

②本研究は解析にデータマイニング手法を導入することが大きな特徴である。その前段階として、データベース（厳密にはデータマート）の構築が必須である。申請者の施設は、外来患者約15,000名、入院58床と世界最大の糖尿病センターである。単一施設なので、データベース化に際して重要なデータの標準化・データのセキュリティーは比較的容易に達成できる。さらに、2003年7月からは、全病院的に電子カルテが導入されており、良質の患者情報を得るためには、最も適した施設である。

データマイニング手法が従来の疫学解析と異なる点は、あらかじめ仮説を設けずに、コンピュータ自身が、決定木、リンク分析、クラスタ分析、ニューラルネットワークなどを通じて、仮説を見出す点にある。特に、糖尿病においては、環境要因が患者のアウトカムやQOLに与える影響について未知な点が多いので、データマイニング手法によってはじめて明らかになる事実も多いものと期待される。比較的少数患者（数千以内）のデータ解析については、ランダム化比較試験、コホート解析などをはじめとした従来の前向きの疫学的解析方法が優れているが、対象患者を限定され、綿密な計画と多額な費用がかかるうらみがある。データマイニング手法はより大規模な蓄積データ（数万件以上）から未知のデータ間の関係を見出すことに主眼を置かれているので、後ろ向きの解析に適しており、研究費用は前向き研究に比べて少なくてすむ利点がある。

また、臨床情報のデータベース化とデータマイニング手法が確立されれば、患者遺伝子の解析研究と統合することにより、特定の遺伝子が疾患にどのように関与している

のか明確になることが予想される。

③糖尿病の分野において、環境要因やQOLを含めた疫学研究は、その重要性が明らかなのにもかかわらず、まだほとんど行われていない。さらに本研究で導入しようとするデータマイニング手法は医学の分野ではまだ黎明期にある。したがって、本研究は最も先駆的な研究となりうる。

研究計画・方法　（平成16年度）

Ⅰ.

データベースの充実

本研究では、糖尿病患者臨床情報として、新たに、食事摂取状況、運動状況、患者の心理的状況、医師・看護師による患者評価、細小血管障害、大血管障害、日常生活機能、社会生活機能を総合したQOLを入力する予定である。これらの情報は多くは、診療録から手入力する。申請者の施設の診療録は一定の形式で書かれており、2003年7月からは電子カルテ化もなされているので、比較的効率よくデータの入力が可能である。それぞれの項目について、最初に定義をはじめとしたデータの標準化を行ってから、入力を開始する。

● 食事摂取情報：初診時、入院時、および定期的な栄養指導情報がファイル化されており、そこからデータを入力する。

● 運動状況、患者心理状況、日常生活機能、社会生活機能：過去の情報については、入院時に聴取された既存の看護師データベースのみが良質な情報源であり、入力対象とする。今後の初診患者については、電子カルテ上に定型的な調査項目を設けたので、そのデータを入力する。

● 細小血管症：当センターは眼科も有しており、受診患者の90％以上の網膜症の定量評価が可能である。腎症は、微量アルブミン、血液生化学検査値より定量的な評価が可能である。神経症は学会全体での評価基準が定まっていないので、初診時と入院時のアキレス腱反射および神経伝導速度のみが、現状では評価可能なデータである。

● エンドポイント：患者の死亡の把握は、従来からシステム化されて行われている。心筋梗塞、脳梗塞、癌の発症などのイベント情報についての過去のデータ収集は、限定的なものに過ぎない。今後、電子カルテに専用欄を設けて、そこから情報を収集する。

データの入力は、菅野宙子、丸山聡子、宇都祐子（研究分担者）が行う。また、データ入力の補助業務を行う検査技師１名に対して、謝金を支払う。

データベースのセキュリティーの確保

現段階では、孤立したサーバと1台の端末PCにのみデータを蓄積しているが、より強固なデータベースを構築するために、本格的なデータベース管理システムとサーバ・クライアントシステムを構築し、研究組織に属する研究者が容易にデータの入出力を行えるようにする。同時に、外部に患者情報が漏れないように、データ管理者[岩本安彦(研究分担者)および大学倫理委員会が指名する者]とデータベース管理者[佐倉　宏（研究代表者）]をおく。

上記システムの構築のために、データベース管理システムソフトウェアを構築する予定である。また、研究分担者用端末用PCを2台購入する。

糖尿病臨床情報の疫学的解析

データベース言語 SQL (Structured Query Language) を用いて、疫学的解析に必要な情報を抽出し、従来の統計学的手法を用いて解析を行う。データマイニング手法を適応する前に、一般的な臨床疫学解析を行うことにより、データベース化した医療情報の全体像を把握することができる。また、データマイニングを行うのに最も適した課題を探る大きな手がかりとなる。

本年度は、すでにある程度解析がなされている、経口血糖降下薬の効果の解析及び初診および入院患者の予後解析、を中心に行う。

① 経口血糖降下薬の効果の解析

経口血糖降下薬の効果の短期的な指標はHbA1Cを代表とする血糖コントロールである。すでに、各種薬物を第一選択薬として用いた時の血糖コントロールの経過、薬物のresponder群、non-responder群に分けた時の背景因子の相違などの検討を行ってきた。薬物開始時の食事摂取情報、運動状況、患者心理状況、日常生活機能、社会生活機能のベースラインデータを入力することにより、患者背景因子と薬物効果の関係がより明確になることが期待される。

② 初診および入院患者の予後解析

初診患者および入院患者は、ベースラインとなる情報がもれなく診療録に記載されているので、疫学的解析に有用である。主として、血糖コントロール、合併症の進展、死亡をエンドポイントとした疫学的な解析を行う。

上記解析は、佐倉　宏（研究代表者）および菅野宙子、丸山聡子、宇都祐子（研究分担者）が行う。

データマイニング手法による解析

データマイニング手法の第一段階として、データの洗浄（はずれ値の検討、入力ミスの排除など）、妥当な導出関数の設置（常識的な例としては、身長と体重からBMI、糖尿病発症年と初診日から糖尿病経過年、を求めることなど）が重要である。

データマイニング効率を上昇させるためにはクラスタ分析による因子の整理が必要である。例えば、本研究では、患者の日常機能評価や社会機能評価に、入院時のアンケートを元にした看護師データベースを利用するが、看護師データベースの多岐の項目は、何種類かの類似項目に分類することができると思われる。このようなクラスタ分析自体についても従来の統計学的解析手法よりデータマイニング手法が威力を発揮する分野である。看護師データベースの項目間のクラスタ分析を行うことにより、患者のパターンの分類が可能になり、日常機能や社会機能評価に繋がるものと期待される。

本年度は、データマイニング解析の有効性を確立するために、主として、上記の疫学解析の項で取り上げた①経口血糖降下薬の効果の解析、②初診および入院患者の予後解析、を平行して行う。データマイニングの中心的な手法である、決定木手法を用いて、経口血糖降下薬の有効性、初診および入院患者の予後の良否を分類指標として、どの因子が重要な役割を果たしているのかを検討する。

統計学的解析およびデータマイニング手法のために、専用のソフトウェア使用の年間契約を行う。また、データマイニング手法の解析者に対して、謝金を払う予定である。

データマイニング手法は、佐倉　宏（研究代表者）が行う。

Ⅱ．生命倫理・安全対策に関する留意事項

医事・薬剤・検査情報に関しては、担当部署に研究計画書を提出して、許可を得てからデータを収集した。今後、電子カルテからの医療情報収集という、二次的な利用も提案しているため、個人情報の保護と倫理的な面には最大限の留意を払う必要がある。そこで、本研究に関して、「疫学研究に関する倫理指針」（平成 14 年 6 月 17 日文部科学省・厚生労働省告示第２号）及び「疫学研究に関する倫理指針の施行等について」（平成 14 年 6 月 17 日文部科学省研究振興局長・厚生労働省大臣官房厚生科学課長連名通知）に基づいた研究計画書を大学倫理委員会に提出して、承認が得られてから研究を進める。

研究計画･方法　（平成17年度）

データベースの充実

平成16年度に引き続き、データベースの充実化を行う。具体的には、薬剤情報、検査情報の電子化が開始された1995年以降に初診し、1年以上の通院歴のある約10,000人の外来初診情報、および約5,000人の入院患者の情報をデータベース化する。

統計学的解析とデータマイニング解析

平成16年度に引き続き、統計学的解析とデータマイニング解析を行う。解析テーマも前年度に引き続き、①経口血糖降下薬の効果の解析および②初診および入院患者の予後解析を中心に行うが、薬物開始時の食事摂取情報、運動状況、患者心理状況、日常生活機能、社会生活機能のベースラインデータの充実とともにより詳細な解析が可能であると予想される。データマイニング手法については、クラスタ分析、決定木手法に加えて、リンク分析、ニューラルネットワーク手法も解析に使用する。

データマイニング手法で得られた仮説の確定

データマイニング手法で得られたデータ間の法則については、コンピュータが探し出してきた仮説であるから、まったく予想外の関係を見出せる可能性を秘めている。しかし、同時に、全く理解不能な関係や最初の解析に用いたデータ（トレーニングセット）にover-fitした仮説を得てしまう危険性も高い。そのため、新たなデータ（テストデータ）を用いて、仮説の検証を行う必要がある。特に、テストデータについては、新患患者や新規入院患者の医療情報を前向きにフォローして得られたデータを用いる。テストデータによっても追認された仮説については、真理である可能性が高い。そして、多くの場合、その新仮説は、従来の統計学的手法を用いても検証することが可能である。

研究成果

データベースの充実

平成 17 年度末までに、約 600 万件の糖尿病医療情報収集を行った。当初予定していた、食事摂取状況、運動状況、患者の心理的状況、医師・看護師による患者評価、細小血管障害、大血管障害、日常生活機能、社会生活機能を総合した QOL は日本糖尿病学会で現在標準化の作業が行われているため、完成した後にデータベースに取り入れることとした。

データベースのセキュリティー確保

サーバ・クライアントシステムに移行して、外部へのデータ流出を防ぐ措置を行った。データベースの構築に関しては既に大学倫理委員会の承認を受けていたが、電子カルテからのデータ移行に関して、さらに病院長へも申請し、承認を得た。

データマイニング手法による解析

本研究の主目的であり、成果を報告する（出版物 1 参照)。

はじめに

情報技術の発達により、膨大なデータ処理が極めて容易なものになった。医療の世界においてもレセプト、検査データからはじまり、オーダリングシステム、電子カルテと急速な勢いで IT 化の波が押し寄せてきている。おそらく 10 年後には、医療機関の診療録はほとんどが電子化されていることであろう。このようにして集積されたデータをうまく解析することができれば多くの有益な知見が得られることは疑いない。しかし、あまりにデータが多いとどのように解析したらよいのかかえって途方に暮れてしまう。

本稿で述べるデータマイニング手法は膨大なデータの中から、コンピュータがデータ間の関連性や規則性を見出し、新知見発見や診療方針決定（decision-making）を支援する新しい手法である。糖尿病の発症・進展・予後・治療などを解析する上で、今後データマイニング手法がおおいに有用になると考えられる。

データベース化に適した糖尿病診療

糖尿病は有病率が高く、また非常に複雑で幅の広い疾患である。つまり

- 患者の予後を左右する合併症は、糖尿病発症後何年も経てから発症する。
- 予後を規定する因子は、血糖コントロール以外にも、血圧・高脂血症・年齢・遺伝・生活習慣など数多く存在する。
- 治療の成否は食事・運動療法という患者自身の実践如何にかかっている。
- 前項の達成には糖尿病患者教育が重要であるが、その方法が確立されておらず、医療施設間の格差が大きい。

以上のように、膨大な数の患者が存在し、多くの因子が疾患の予後に影響を及ぼす複雑な疾患を理解するためには、医療情報をデータベース化し、多角的な視点から解析していく必要がある。

データマート（データウェアハウス）の構築（図 1）

糖尿病情報のデータベース化およびその解析を行うにあたり、まずデータの定義・標準化を行って、各部署にばらばらに存在するデータを統合する必要がある。例えば、生年月日、初診日、性などの基本情報は医事課、検査情報は検査部、処方情報は薬剤部が管理している。これらはすでに電子化されていることが多いので、技術的には容易に統合できる。しかし、身長・体重・血圧・自覚症状などは診療録に手書きされていることが多いため、情報の統合化は簡単ではない。また、糖尿病の診療において、遺伝子、心理・行動、社会・経済、家族、医療側の体制なども明らかに重要であるが、これらは診療録にすら書かれていないのが実情であろう。今後、これらの情報をいかに定義して、データベース化するかは大きな課題である。このように、散在している情報を統合したデータベースをデータマート（データウェアハウス）と呼ぶ。データマートは Excel のような表の集合体であり、それらがキーで関連づけられて、リレーショナルデータベースという構造をとっている。

データマート解析の 5 段階（図 2）

データマートはその膨大な情報を解析し、意思の決定（医療データマートにおいては新しい診療指針の提示）に結びついてはじめて価値があるものになる。データマートの解析には次の 5 段階がある。第 1 段階はデータマートに統合された膨大なデータの中からデータベース言語 SQL (Structured Query Language) を用いて必要な情報の抽出を行う操作である。SQL 言語の知識がなくてもグラフィカルなインターフェイスによって情報の抽出を支援するプログラムも存在する。第 2 段階は抽出した情報を統計学的手法で解析することである。SQL で抽出した情報は通常 Excel のような表形式であるので、一般的な統計ソフトを使って容易にデータ解析を行うことができる。第 3 段階は、多角的視点からデータのサマリーを瞬時に提示する OLAP (On-Line Analytical Processing) 機能である。この段階までは、あらかじめ想定した仮説に沿ってデータの解析を行い、仮説が正しいか否か検証を行うステップといえる。

本稿で述べるデータマイニングは第 4 段階にあたる手法であり、これはコンピュータがデータ間の関連性や規則性を見出し、新知見発見や診療方針決定（decision-making）を支援する、つまりコンピュータが仮説を発見という点で、第 1－3 段階と全く異なっている。第 5 段階は最適化解析とも呼ばれているが、これはもっともすぐれた仮説を見出す機能である。

データマイニングの実例

データマイニングについて、実際の解析例を提示しながら説明する。

血糖コントロール目的で入院した患者には、コントロールが良好になる人とならない人がいるが、その違いはどのような因子が関与するのか検討することにした。東京女子医科大学糖尿病センターに 2000 年 4 月から 2003 年 3 月に入院し、入院時の HbA1c 値が 8%以上で、重大な併発疾患・合併症のない 676 名を対象とした。すでに構築してあるデータマートおよび診療録から、入院時の年齢・BMI・HbA1c、糖尿病型、中断歴の有無、退院時治療、

診断～入院、初診～入院、入院回数を抽出した。入院6ヶ月後の血糖コントロールが7%未満をコントロール良好、7%以上をコントロール不良とした。

データマイニングソフトウェアとして、研究目的になら自由にdownloadすることのできるWEKAプログラムを用いた。このプログラムはアルゴリズムが豊富でかつ教科書も存在するので、非常に使いやすいソフトウェアである。表形式のデータはWEKAに簡単に入力することができ、すぐにデータマイニングを行うことができる（図3）。ひとくちにデータマイニングといってもさまざまな手法があり、それぞれについても何種類ものアルゴリズムが開発されている。WEKAでは、いったん入力したデータについて、さまざまなアルゴリズムを用いて解析することができる（図4）。

- **属性選択(Attribute Selection)**（図5）

抽出した因子の中でどれが6ヵ月後の血糖コントロールに重要な因子であるかをさぐるために、Attribute Selectionを用いて解析したところ、図5のように、診断から入院までの期間、初診から入院までの期間、入院回数、退院時治療法、糖尿病型、中断歴の純となり、入院時HbA1c・BMI・年齢は6ヵ月後のコントロールとは無関係であることがわかった。同様の解析は、多変量解析を用いても行うことができ、実際ほとんど同じ結果を得た。一般にデータマイニングでは、1症例について、多くの因子（属性）を入力するので、Attribute Selectionによって、重要な属性を選択しておけば、後の解析に有用であると言われている。

- **アソシエーションルール（Association Rule）**（図6）

各因子間の関連性について、条件式で記述したものである。図5の1や19については、当然過ぎて有用な情報とは言えないが、104、152、156、164、205などは面白い情報と言えるので、さらに詳細な検討をする価値があるだろう。

- **決定木（Decision Tree）**（図7）

上から順に条件分岐をたどっていくことにより、最終HbA1cが7以上になるか7未満になるかわかるので、非常に理解しやすく、データマイニングの中でももっともよく行われるアルゴリズムである。

さまざまに存在するデータマイニング手法の中で、どれを用いたらもっとも良いかは、実際の解析例ごとに異なるだろう。しかし、どの方法を用いたとしても、データマイニングによって発見した仮説を、さらに詳細に検証する必要がある。

データマイニング手法が有望と考えられる糖尿病の分野

データマイニング手法は、HbA1c、血圧、コレステロール、BMIなどすでに重要性が確立されている指標よりも、重要そうではあるがまだ確立されていない指標を用いた解析、膨大なデータに基づいた後ろ向き(retrospective)な解析に向いている。下記の糖尿病の分野は特にデータマイニング手法が有用と考えられる。

- 遺伝子解析…データマイニング手法を用いた論文も出てきている。
- 薬物効果解析…経口血糖降下薬の効果は食事・運動療法に大きく影響を受ける
- 患者教育…指標の定義・数値化が困難で、統計学的な解析が困難
- クリニカルパス

おわりに

本稿では述べなかったが、データベース化とその解析にあたってはする上で、個人情報の保護と「疫学に関する指針」の遵守が重要であることはいうまでもない。

また、日本全体の糖尿病診療の実態を把握するためにデータベースを構築する重要性が日本糖尿病学会でも言われ始めているが、複数の施設のデータ比較などを行う上では、データの標準化が重要である。今後、日本糖尿病学会にデータ標準化委員会が設置されることが望まれる。

図1

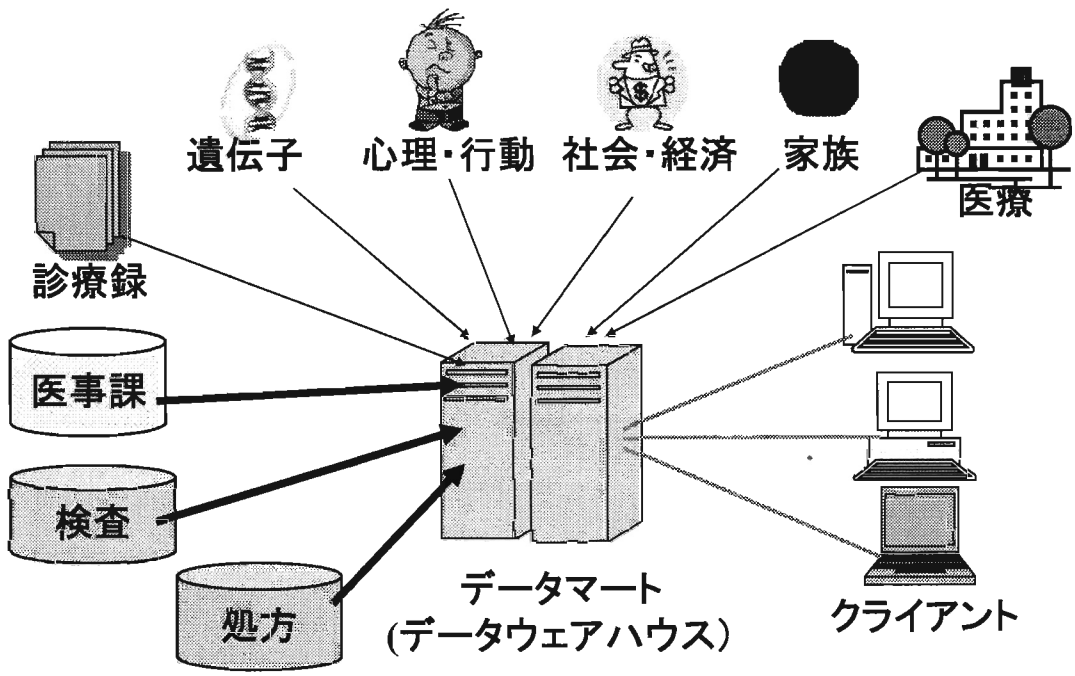

図 2

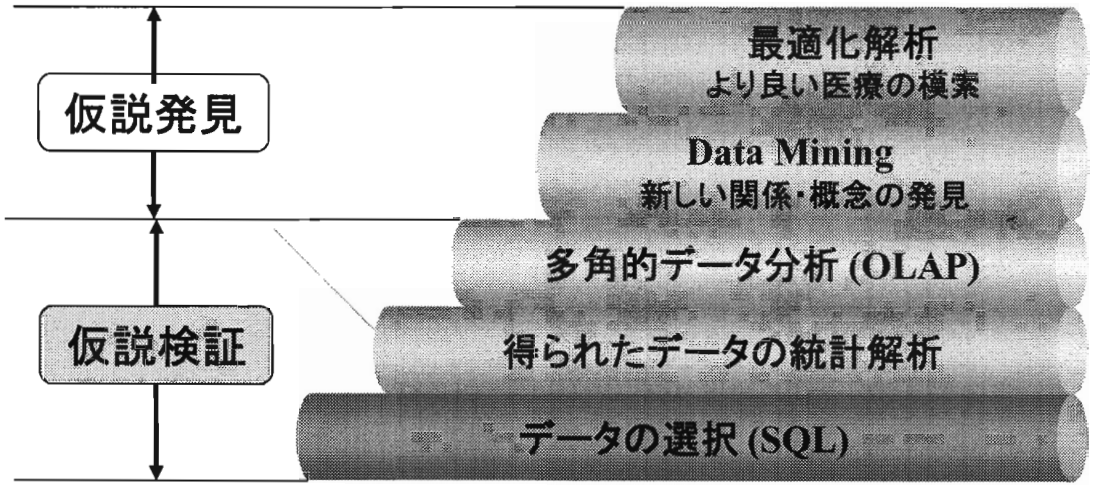
仮説発見
仮説検証
最適化解析
より良い医療の模索
Data Mining
新しい関係・概念の発見
多角的データ分析 (OLAP)
得られたデータの統計解析
データの選択 (SQL)

図 3

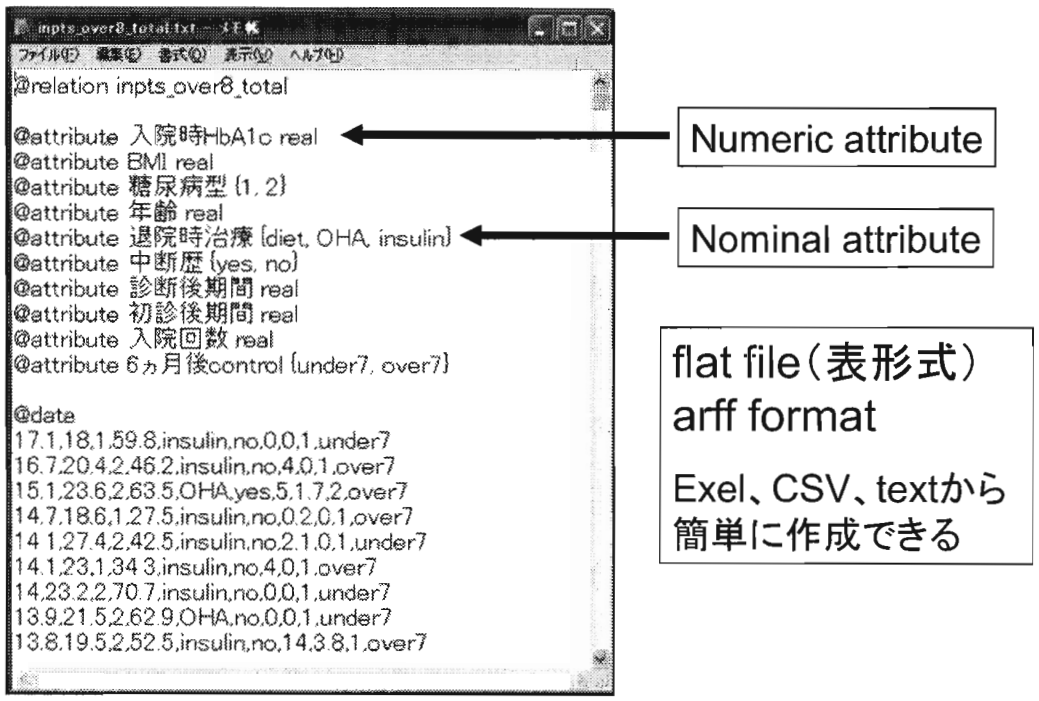
@relation inpts_over8_total
@attribute 入院時HbA1c real
@attribute BMI real
@attribute 糖尿病型 {1, 2}
@attribute 年齢 real
@attribute 退院時治療 {diet, OHA, insulin}
@attribute 中断歴 {yes, no}
@attribute 診断後期間 real
@attribute 初診後期間 real
@attribute 入院回数 real
@attribute 6ヵ月後control {under7, over7}
@data
17.1,18,1.59.8,insulin,no,0,0,1,under7
16.7,20.4,2,46.2,insulin,no,4,0,1,over7
15.1,23.6,2,63.5,OHA,yes,5,1.7,2,over7
14.7,18.6,1,27.5,insulin,no,0.2,0,1,over7
14 1,27.4,2,42.5,insulin,no,2,1,0,1,under7
14.1,23,1,34 3,insulin,no,4,0,1,over7
14,23.2,2,70.7,insulin,no,0,0,1,under7
13.9,21.5,2,62 9,OHA,no,0,0,1,under7
13.8,19.5,2,52.5,insulin,no,14,3 8,1,over7
Numeric attribute
Nominal attribute
flat file（表形式）
arff format
Exel、CSV、textから
簡単に作成できる

図 4、

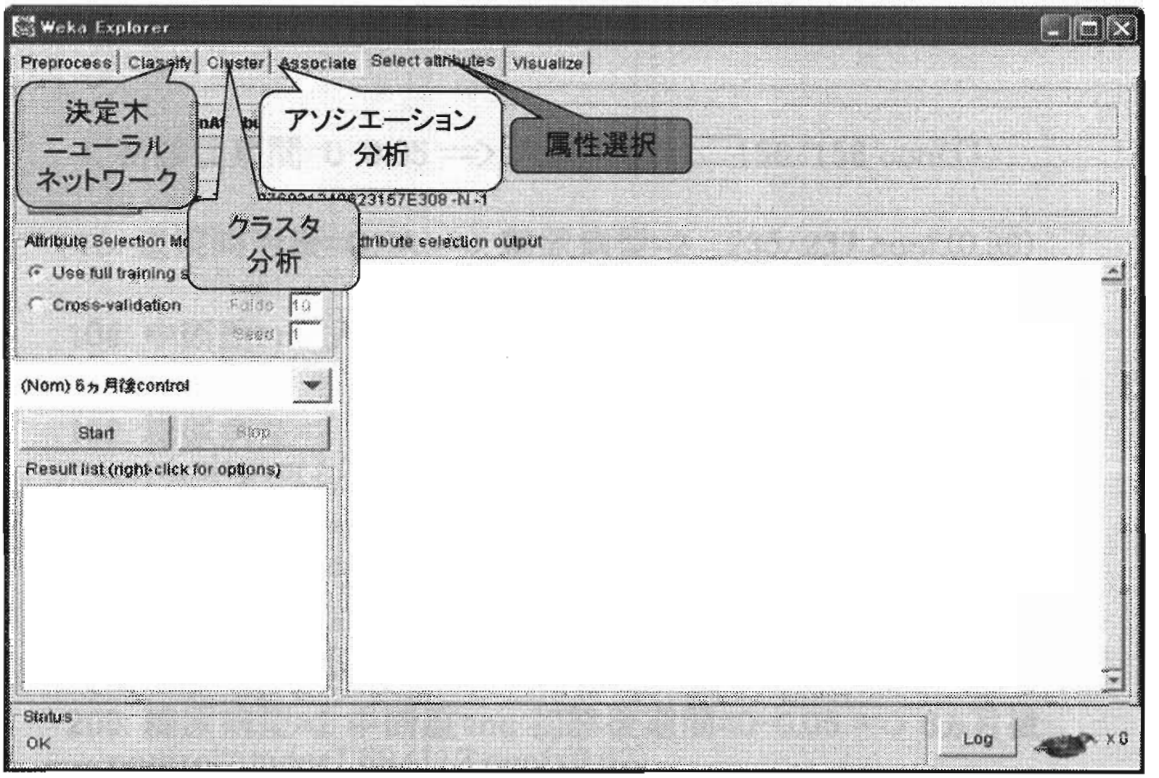

図 5、Attribute ranking

Search Method: Attribute ranking

Attribute Evaluator (supervised, Class (nominal): 6カ月後control):
Information Gain Ranking Filter

Ranked attributes:

0.09224	診断～入院期間
0.073978	初診～入院期間
0.05863	入院回数
0.017992	退院時治療
0.001612	糖尿病型
0.000241	中断歴
0	入院時HbA1c
0	BMI
0	年齢

図6、Association Rules

1. 初診後期間=0-0.05 ==> 入院回数=1　156/156 conf:(1)

⋮

19. 退院時治療=OHA ==> 糖尿病型=2　235/237 conf:(0.99)

⋮

104. 中断歴=no 入院回数=2 ==> 6カ月後control=over7 70/88 conf:(0.8)

152. 糖尿病型=2 退院時治療=insulin==> 6カ月後 control=over7 227/317 conf:(0.72)

156. 退院時治療=insulin 中断歴=no ==> 6カ月後control=over7 218/308　conf:(0.71)

164. 退院時治療=OHA 中断歴=no ==> 6カ月後control=over7 125/179 conf:(0.7)

205. 糖尿病型=2 中断歴=no 初診後期間=0-0.05 ==> 6カ月後 control=under7 69/112 conf:(0.62)

図7、Decision Trees