論文

生活習慣病予防教室における効果傾向抽出のための データマイニングシステム

A System to Mine Hidden Effects from the Classes for Preventing Lifestyle Related Diseases

吉廣卓哉 [1*]，井上悦子 [1]，田部浩子 [2]，中川優 [1]

Takuya YOSHIHIRO, Etsuko INOUE, Hiroko TABE, and Masaru NAKAGAWA

1 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University
〒640-8464 和歌山市栄谷 930 番地
E-mail: {tac, etsuko, nakagawa}@sys.wakayama-u.ac.jp
2 和歌山大学システム工学研究科
Graduate School of Systems Engineering, Wakayama University
〒640-8464 和歌山市栄谷 930 番地

本研究では，生活習慣病予防教室により得られる身体測定や血液検査等のデータから，生活習慣病の予防効果の傾向を把握するための分析支援システムを開発した．近年，全国的に生活習慣病予防教室が実施され，予防への取り組みが活発化するに従い，どのような取り組みをすればどのような参加者に効果があるのか，その詳細を分析することの重要性が高まっている．本システムでは，教室を実施することで得られる時系列の測定データから，どのような群の参加者に効果があったのかを組み合わせ的に調べて分析者に提示することで，少ない労力で教室の効果の傾向を把握できるようにした．また，分析者は，提示された群のデータを表形式で表示して分析結果を確認した上で，群の切り出し条件を少し変えた場合等の周辺結果を試行錯誤しながら閲覧でき，本システムを分析および現場での取組み改善のための基本的なツールとして使用できる．

In this paper, we present a system to support medical doctors to derive hidden effects of the classes that are organized to improve lifestyle related disease. Recently in Japan, the number of such kind of classes is increasing rapidly, but it is not easy to know the type of participants for whom each class has particular effects. Our system does a data mining to find a group for which a class has particular effect by generating several combinations of conditions from the data of the classes. With our system, one can find groups of particular effect without much time and labor, and can also confirm the original data from which the effect is revealed. Our system can be used as a fundamental analysis tool to analyze the data of that kind of classes.

キーワード: 生活習慣病，データベース，データマイニング，Web アプリケーション
Lifestyle Related Diseases, Database, Data Mining, Web Application

1. はじめに

近年, 世界的に生活習慣病への問題意識が高まっている. 生活習慣病とは, 食習慣や運動習慣などの生活習慣が原因となる疾患群の総称であり, 代表的なものとして糖尿病や高血圧, 高脂血病などが挙げられる. 生活習慣病の共通要因として肥満が挙げられることが多く, このためメタボリック症候群[1]とも深い関係がある. 厚生省の調査[2]によると, メタボリックシンドロームは主に中高年層に多く, 40 歳～74 歳では予備軍を併せると男性では 2 人に 1 人, 女性では 5 人に 1 人が該当する. 生活習慣病の一つで, 肥満とも関係の深い糖尿病に関しては, 平成 14 年の糖尿病実態調査[3]によると, 日本では予備軍を入れて約 1600 万人が該当し, 一方, アメリカ政府の調査[4]では, アメリカ人のうち約 2000 万人が該当するとの報告がある. いずれの症状も継続的な増加傾向にあり, 社会的にも大きな問題として認知されるようになっている.

近年, このような生活習慣病の状況への対策として, 全国的に生活習慣病予防教室の実施が活発化している. 生活習慣病改善のためには, 食事の改善と運動の促進が有効であるとされ, 食事・運動療法の知識を増進する講義や, 効果的な食事の改善法や運動方法の実技指導などが, 全国の市町村や病院を中心として行われている. しかし, これらの取組みは日が浅い場合が多くノウハウが十分に蓄積されておらず, 保健師などの指導者が経験や勘に基づいて試行錯誤で行っていることが多い. 保健師の人手不足などもあいまって, より効率的かつ効果的な教室の実施が求められている.

一方, 生活習慣病予防教室では一般的に, 参加者は複数回にわたって身体測定や血液・尿検査, アンケート調査などを行う. 2008 年 4 月からの特定検診の義務化の影響もあり, 同様の教室が全国的に展開されており, また特定検診でも類似したデータが得られるため, そのデータ量は膨大になると予測される. この膨大なデータを用いた分析, およびデータマイニングへの取組みも始まっている. 文献[5]は, 蓄積された膨大な生活習慣病データとその分析結果を利用して, 患者に対する生活習慣病リスクの判定を実現している. また, 生活習慣病データからのデータマイニングにより, 須賀ら[6]が睡眠時無呼吸症候群と高血圧発症の関係性を導き出すなどの成果も挙がっている.

また, 個人の生活習慣病予防に目を向けると, 個人の毎日の健康管理データを用いた分析も試みられている. 竹内ら[7][8]は個人の毎日の健康データをシステムに入力することで, 自動的に相関ルールを発見しユーザに提示できる個人健康管理システムを開発し, 健康管理にデータマイニングを活用することを提案している.

しかし, 生活習慣病データが発生する教室実施現場では, このデータを利用して現場の教室実施の効率化につなげるための分析は十分ではない. 一回の教室で発生するデータ項目が膨大であることに加え, Excel を含めて一般的な統計解析ソフトだけでは, 時系列データを含む教室データから, 有用な傾向を発見することは難しいからである. ここで, 本研究で扱う時系列データは, 個人が検診等を受診した時に発生するデータであり, 連続的に大量の値を含むものではないため, 統計解析ソフトでも必ずしも取り扱いができないわけではない. しかし, 個人の健康管理データから有用な傾向を読み取る分析作業には, 項目の組合せ数が多いことから大きな労力がかかり, これを現場で統計解析ソフトにより行うことは現実的ではない. また, 統計解析ソフトによってはスクリプトを用いて処理を自動化することができるが, 現場でこのような高度なプログラムの利用を想定することも現実的ではない.

そこで本研究では, まず生活習慣病予

防教室から発生するデータをモデル化し,多くの教室データを統一的に扱えるようにデータベース化した．次に，蓄積されたデータから,どのような条件の参加者に教室の効果が見られたのかという傾向を把握するための,データマイニング手法を提案し,これをマイニング支援システムとして実現した．さらに，実際の生活習慣病予防教室のデータを用いて評価を行った．

本論文の構成は以下の通りである．まず，2 章で生活習慣病データについて説明したのち,データから効率的に傾向を把握するデータマイニング手法について述べる．3 章では実装したシステムの動作及びインタフェースについて説明する．4 章で評価と考察について述べた後，5 章でまとめる．

2. 生活習慣病の予防効果分析

2.1 生活習慣病の予防効果分析

生活習慣病とは，食習慣，運動習慣，休養，喫煙，飲酒などの生活習慣が，その発症･進展に関与する疾患群の総称であり，糖尿病，高血圧，高脂血症，肥満症,高尿酸血症などの病気が挙げられる．これらの病気は,生活習慣を改善することで,発症や進行が予防できるということが明らかになっており,全国的に予防への取り組みが急激に活発化している．予防方法としては継続的な運動や食事の改善が主であるが,いずれも一定の知識のもとに,忍耐強い個人の取り組みが必要であり，知識獲得のための講義や,効果的な運動・食事の改善方法の指導,そして個人が意識的に予防活動を継続できる仕組みつくりが求められている．

このため,全国の病院や市町村を中心として,運動や食事の改善を主とした生活習慣病予防が実施されており,栄養指導や生活習慣病の講義,効果的な運動の指導など様々な取り組みが行われている．これらの教室では一般的に，教室実施による効果や傾向を(参加者と指導者の両方が) 把握するために，教室の実施期間中に数回にわたって,複数の調査や測定が行われる．この値を用いて，参加者と指導者は教室による効果を確認することになる．

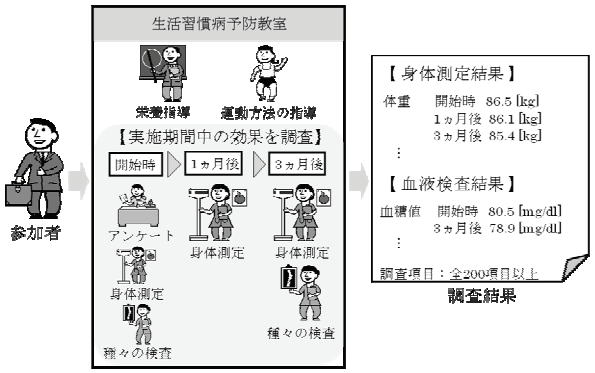

図 1. 生活習慣病予防教室の概要とデータ

このような生活習慣病教室は近年継続して取組みが行われる場合も増加し,保健師らによる経験も蓄積しつつある．また,今後もこれらの活動は全国的に広まるものと推測される．

2.2 予防効果分析へのニーズ

一方,生活習慣病教室への取り組みが継続するにつれて,どのような取組みがどのような参加者に対して効果があるのかを把握することで,効率的に教室を運営したいという要求が発生してきている．年齢層や性別，体質，地域，生活習慣のパターンなどで参加者を分類し,どのような参加者に対してどの取組みが有効であるかを把握できれば,より効率的な教室の運営が可能になることは言うまでもない．しかし現状では，このような傾向把握は専ら教室を担当する指導者が積み上げたノウハウという形で累積されており,体系だった知識が構築されているとは言い難い状況である．

一方,蓄積された教室データを用いた分析への取組みについては，分析者が Excel などの一般的な統計ソフトを用いて個人で分析を行う場合が多く,その場

合には手間の問題から，その教室の主催者が工夫をした点について効果を検証する程度の分析しかなされず，網羅的な分析がなされないままデータが放置されることになる．しかし，このデータを用いて隠れた傾向を読み取ることは，教室実施を効率化するために重要なことである．本研究ではこのニーズを汲み，手間なく網羅的な分析ができるシステムを実現する．

教室を実施することで得られる情報は，教室の参加者名簿と参加者の検査・調査結果である．教室は通常，一定の期間内に数回参加者に集まってもらい，指導と調査（検査）を行う．調査は，アンケートや身体測定，血液検査，尿検査などに分類され，教室により，どのタイミングにどの調査を行うかが異なる．もちろん，同じ調査（例えば身体測定）が違う時期に複数回実施される場合もあるため，調査データは時系列データとなる．教室によって検査・調査項目は多少異なるものの，教室には複数の参加者が参加し，参加者ごとに時系列での調査結果が存在するというデータ構造は同じである．図 1 に生活習慣病予防教室の流れとデータの概要を示す．この図には，参加者一人一人に対して複数回にわたり各種の調査や測定が行われるデータ構造が示されている．

このデータから有用な傾向を読み取るにはどのような分析をすれば良いだろうか．求められていることは，各教室に対して，どのような条件の参加者の群に対してとりわけ効果があったのかを知ることである．教室は幾つもの地域で実施されており，教室ごとに生活習慣病予防を目的とした特色ある手法による取り組みが行われているため，それぞれの取り組みが参加者に及ぼす効果傾向は少しずつ異なると考えられる．また同様に，教室の中でも参加者によって得られる効果には違いがあり，それにも何らかの理由があると考えられる．実際，教室には年齢や性別の異なる参加者が存在しているように，その他にも喫煙者や非喫煙者，酒を飲む人や飲まない人，太っている人や痩せている人など，個々の参加者が異なる条件を持っており，参加者の持っている条件が効果の違いに大きな影響を及ぼしている可能性は十分に考えられる．

そこで，教室を実施して得られた調査項目の値を用いて様々な条件で参加者をグループ化し，明らかな効果の違いが表れる群を見つけ出すことを試みる．これにより，どのような条件の参加者に効果があったのかを知ることができる．

2.3 予防効果分析のためのデータ比較

どのような条件の参加者にどのような効果があったかを把握するために様々な分析方法が考えられるが，本研究のアイデアは次のようなものである．まず，効果を上げるとは，ある一つの調査項目の値が，教室の取組みの経過と共に統計的に有意に変化することである．例えば，ある参加者の群について，教室の経過に伴い体重の減少に有意な差が見られれば，この教室は体重を減らす効果があったと言える．さらに，ある条件を満たす参加者に対して効果があることとは，ある条件を満たす参加者の体重差には有意差が認められるが，満たさない参加者の体重差に統計的な有意差が認められないことである．実際には，それほど極端に有意差の有無が分かれることは少ないかもしれない．しかしこの傾向の程度は，次のような比較を行うことによって測ることができる．

A) 参加者をある条件により 2 群に分割する．分割した各群に対して，ある調査項目の 2 つの時期の値の差を（参加者ごとに）とり，その差の程度を比較する．

つまり，一つの教室の参加者に対して，様々な条件を用いて上記 A の比較を行うことが傾向把握につながる．なお，はじめに教室の参加者を何らかの条件(男性のみ，など)で絞り込んでおいて，そ

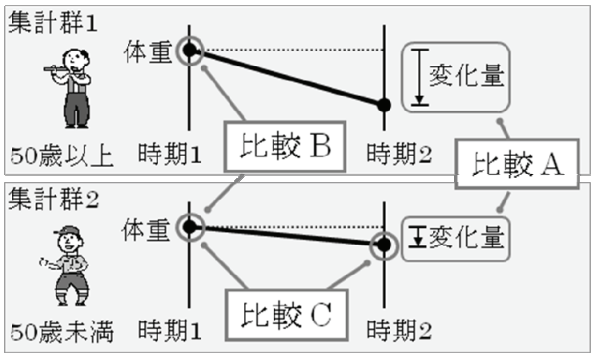

図 2. 三種類の群間比較

の後に比較 A を試しても良い.

例えば，40 歳以上の参加者を体重が 60kg を超えるか否かという条件で 2 群に分割し，教室開始時と 3 ヵ月後の BMI(Body Mass Index, BMI=体重(kg)÷身長$(m)^2$)の差を比較する．この「差」に統計的に有意な差が見られれば, 体重によって BMI の変化傾向に差があることが読み取れる．さらに，比較 A に加えて次の比較 B, C を補助的に用いると, 傾向をより正確に把握できる.

B) 参加者をある条件により 2 群に分割する．各群の参加者に対してある時期のある調査項目の値を 1 つとり，この値を群間比較する.
C) 参加者をある条件により絞り込むことで 1 つの群を得る．この群に対して，ある調査項目の 2 つの時期の値の差を（参加者ごとに）とり，その差の程度を比較する.

比較 B と C は，比較 A の補助的役割を担う．つまり，比較 A によって，ある分割条件に対して何らかの効果の差が見られたとしても, もともとの値に差がどの程度あったのかによって, その効果に対する評価は変わってくる. 例えば前述の BMI の例では，はじめからかなり体重差のある二群の参加者間に明らかな BMI の差があったとすれば（比較 B により判断可能），体重の重い群の参加者により大きな効果が見られても, 驚く傾向とは言えないだろう．また，たとえそうであったとしても, 体重の軽い群の BMI がほとんど変化せず，体重の重い群に限定して BMI の減少が見られたならば（比較 C により判断可能），これは何らかの興味深い傾向を示しているとも判断できる．このように，傾向を正しく把握するには比較 B, C が必要である. 図 2 はこれら三種の比較の関係を図示したものであり, 年齢によって分割した二群の体重の変化を比較している例である. いずれの比較も値の変化を把握するために重要な値の差を表すことがわかる.

これら三種類の比較結果は，t 検定[9]の結果得られる p 値を用いて評価できる．即ち，p 値が低いほど，分割した二群の値に明確な差があると結論できる. 但し, 有意な差があると結論するためには, 有意水準に基づいた一定値を下回ることは必要条件である. 再度図 2 を見るとわかるが，群の分割条件，比較する調査項目, そして調査時期二つが決まれば, 比較 A に対しては p 値が一つ, 比較 B, C に対しては二つずつ定まる. そこで本研究では，分割条件と調査項目，調査時期を選ぶことで比較 A, B, C の結果を一覧できる分析インタフェースを持つシステムを設計・開発した.

2.4 効果傾向のマイニング

分析者は様々な条件で比較 A, B, C を行うことで効果傾向を分析する. しかし，1 教室あたり最大で 200 以上の調査項目があり, それぞれの測定値が時系列上に複数存在するため, その比較の組合せが膨大になり, その全てを手作業で調べることは事実上不可能である. そこで, この膨大な組合せから興味深い条件を効率よく見つけ出せるような, データマイニング機能を提供する.

データマイニングは, 一つの教室に対して分割条件，調査項目，二つの調査時期を組合せ的に網羅し, 各組合せに対して比較 A, B, C を適用することで行う. 比較 A, B, C のそれぞれに対して，p 値

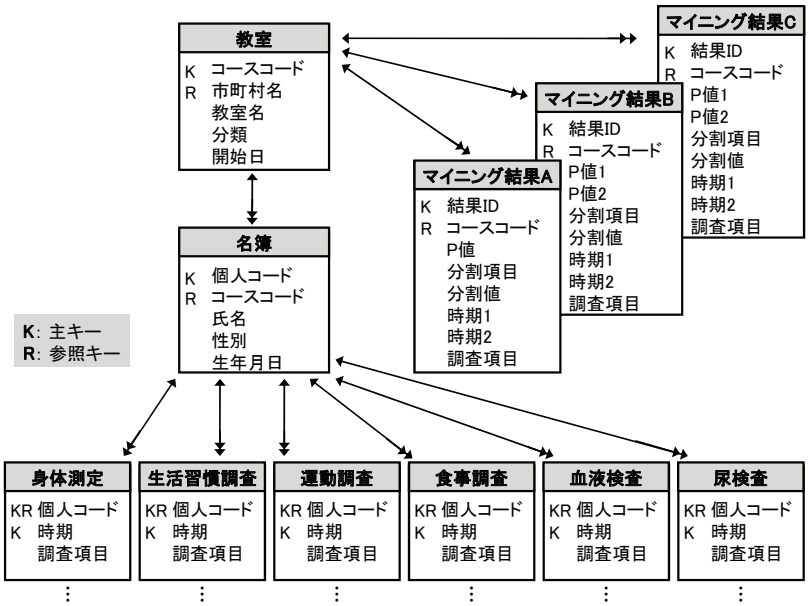

図 3. データベーススキーマ

が低い順に組合せを列挙することで, 分析者は傾向が明確に現れている順に比較結果を確認し, 真に興味深い傾向を効率的に探すことができる.

分割条件の組合せは, 次のように列挙する. 分割に使う調査項目が性別や喫煙の有無のようなカテゴリデータであれば, その値により参加者を二つに分割する. 本システムの調査項目には 3 値以上の値をとるカテゴリデータは存在しない. 分割に使う調査項目が連続数値データ, 例えば体重の場合は, 参加者数を五分割する 4 つの境界値を個別に用いて参加者をそれぞれ二分し, その 4 種類の分割条件を列挙する. 分割数は必ずしも五分割でなくても良いが, 今回は教室あたりの参加者数がおおよそ数十名であることを想定し, 計算量を考慮した結果五分割とした. 比較したい調査項目としてカテゴリデータは使用せず, 全ての連続数値データ項目を列挙する. 調査時期はデータの存在する調査時期二つの組み合わせを全て列挙する.

このように列挙された全組合せに対する比較 A, B, C の結果を計算し, データベースに格納しておく. なお, 群に分割した際に該当データ件数が一定に満たないもの, 例えば片方の群のデータ件数が全体の二割にも満たない場合には, 信頼性が低いとして破棄することとした.

3. 生活習慣病予防教室の効果分析支援システム

3.1 教室データのデータベース化

教室実施により得られる教室の参加者名簿, 各参加者の検査・調査結果などの情報を関連付けて管理し, データ比較や効果傾向のマイニングを行うために, 教室データをデータベース化した. どの教室においても複数の参加者が存在し, 参加者ごとに時系列の調査結果が存在するというデータの関係になっているため, 図 3 に示すように教室を根として木状にテーブルが関係するスキーマ設計とした.

図 3 に示すとおり, データベースは 11 個のテーブルから構成される. 図の K は主キー, R は参照キーを示している. 図の中央上部にある教室テーブルは各教室の基本的な情報を管理するものであり, 各教室に一意に割り振られたコースコードを主キーとする. 教室の下には

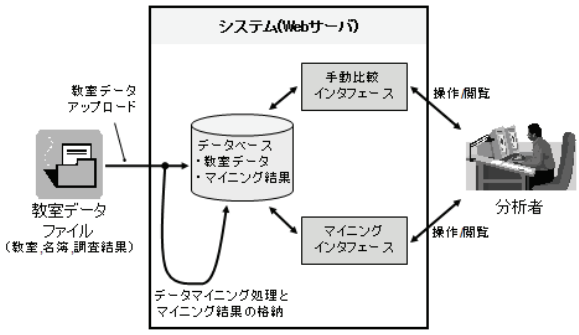

図 4. システム構成

その教室の参加者を管理する名簿テーブルがあり, 各参加者の調査結果データは生活習慣調査, 運動調査, 食事調査(これら三つはアンケート調査), 身体測定結果, 血液検査結果, 尿検査結果の六つの調査種別に分類され, 対応するテーブルに格納される. これらの調査項目は, 基本的には特定保健指導において調査・測定される調査項目を用いており, 一般的にはどの教室にも共通して適用できると考えられる. 但し, 一部には教室独自の取り組みにより新たな調査項目が必要となる場合があり, その場合にはスキーマへの調査項目の追加が必要になる.

調査結果は得られた時期のデータと共にテーブルに格納される. ここで本システムでは, 時期は教室開始時は"0", 3 か月後なら"30"という具合に対応する数値で表現することとした. これは, 教室によっては 3 か月後の測定を(参加者の都合や測定項目によって) 2 回に分ける, 或いは半月程度の間をあけて測定をするなど, 特殊な運用をすることがあるためであり, システムを運用に柔軟に対応させるための仕様である. 数値と時期との対応は, システム利用者が取り決め, 利用者が複数人いる場合には合意しておけばよい.

データマイニングの結果を格納するための三つのテーブルは, あらかじめ全ての組合せに対して実行した比較結果を格納するものであり, A, B, C のいずれの比較に対しても, 一回の比較結果を 1 レコードで格納する. これは, t 検定による比較結果を計算するために相応の時間がかかるからであり, このように予め計算しておいた値を閲覧時に検索・表示する仕組みにすることで, 高速な結果閲覧が可能になる.

また, 本システムは, 一つの組織内で実施される教室データを管理するため, 数十から百数十程度の教室数を想定するが, これを超える場合であっても, インデックスの作成や, 必要に応じて分散データベース化する等の方策によって対処すれば高速な閲覧が可能になると考えられる.

3.2 システム構成

提案するマイニング手法に基づく Web システムを構築した. 本システムの構成を図 5 に示す. 本システムは Web サーバとして構築されている. システムの実装には, DBMS として PostgreSQL 7.4.13 を, Web サーバソフトウェアとして Apache 2.0 を用いており, また, ユーザインタフェース部分は PHP 4.3.9 を用いて構築した.

ユーザはまず, 教室データ (教室自体の情報, 参加者名簿および調査結果) を指定された Web ページからアップロードする. ユーザは指定されたテンプレートを用いて教室データを作成し, これを CSV 形式で保存したものをアップロードする. アップロードを行うと, 教室データがデータベースに登録されると同時に, データマイニングプログラムが動作を開始し, 教室データから傾向を抽出する. 傾向の抽出が終了するとマイニング結果がデータベースに格納される. 上記のようにデータをシステムに登録した後, ユーザは Web ブラウザで手動比較インタフェースとマイニングインタフェースにアクセスし, 教室の傾向把握作業を行う.

なお, 本システムは, 生活習慣病教室を主宰する組織において, 実施される教室データを経時的に蓄積し, その教室データを一元管理し, 担当者間で共有する

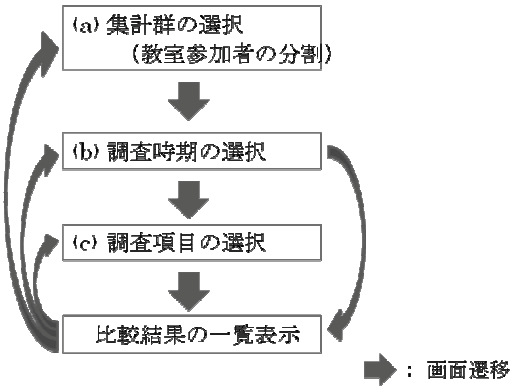

図 5. データ手動比較の手順

図 6. 集計群選択画面

利用方法を想定している．従って，データ量として，少なくとも 100 人程度の参加者が 5〜6 回の調査・測定を行う教室を数十程度は管理できる必要がある．もしこれを超えるデータ量を扱い，サーバの性能が不足する場合には，インデックスを設定する等のチューニングによりパフォーマンスを向上する必要がある．

Web アプリケーションとして実現することで，利用者がソフトウェアをインストールする手間を省くことができ，また利用者の計算機環境に依存することなくデータを集中管理できる．

3.3 手動比較インタフェース

データベース中のデータに対して効率的な比較作業を行うインタフェースについて説明する．効果傾向の分析にあたっては，まず効果傾向を様々な視点から詳細に調べられることが必須である．本インタフェースでは，手動で分割条件や比較項目を入力することで，A, B, C の各比較結果を詳細に確認することができる．また，分割条件等を少しずつ変更して試行錯誤することで，周辺データの傾向を含めて把握する分析作業を行える．この試行錯誤は，後述のマイニングインタフェースを用いて抽出した効果傾向を，さらに詳細に分析し把握するために特に有効である．

手動で A, B, C の比較を行うためには，分析者はいくつか指定すべき項目がある．それは，(a)比較したい集計群の決定，(b)比較したいデータの調査時期（図 2 のように通常は 2 つの調査時期を指定する），(c)比較したい調査項目，の三つである．本システムでこれらの項目を指定する手順を図 5 に示す．分析者は(a)(b)(c)に対応する各画面で表示したい比較に関する項目を指定することで，最終的に比較結として果を画面に表示できる．また，結果表示後に少し比較条件を変えて試行錯誤しながら知識を得られるように，再び各項目の指定画面に戻って一部の項目を変更した比較結果を素早く表示することができる．

以下，(a)(b)(c)の各指定画面について説明する．図 6 は教室の参加者を複数の群に分ける条件を指定する集計群選択画面である．この画面では，上部の「絞り込み検索」インタフェースで教室を検索したうえで，その下の「分割条件」インタフェースで条件を指定して参加者を分割できる．図 6 は体重 60kg を閾値

図 7. 調査時期選択画面

として各教室の参加者を二群に分割した例である．下部の一覧表により，分割された各群の人数がわかり，分割バランスが適切であるかを判断できる．また，この表の右部のチェックボックスにより，分割されたどの群を，比較したいどの集計群に入れるかを選択できる．本システムでは，この画面で作成した集計群同士を比較することになるが，このインタフェースでは手動比較をより柔軟にするために，異なる教室の群を同じ集計群に入れることが可能であり，また 3 つの集計群に分けることができるため，単に条件により二分割するよりも柔軟な集計群の作成が可能である．

図 7 は比較する二つの調査時期を選択する比較時期選択画面である．この画面の上部には，集計群選択画面で選択された群に関する情報が表示され，下部には各集計群がどの調査時期にどの調査種別の調査データを何人分持つかを表す一覧表である．つまり分析者は，これらの情報を参照しながら比較したい調査時期を選択することができる．調査時期はシステムでは数値（図 7 では 0, 1, 20）として表現されているが，これは例えば 0 は事前調査，1 は教室開始時，20 は二か月後の測定などと，データ入力時に数値に意味を割り当てて使用している．なお，調査時期の選択は画面左下部のチェックボックスで行い，三つ以上の調査時期を同時に選択することも可能である．後述のデータマイニング手法では二つの調査時期に限定して処理を行うが，手動の比較作業ではできるだけ多くの調査時期を同時に比較する方が効率良く分析できるため，このようにした．

ステップ(c)に対応する表示項目選択画面は，比較したい調査項目を選択する画面である．調査項目は 6 つの種別に分類され，その項目数は多いものの，この画面は単にチェックボックスを用いて比較したい調査項目を選択するだけなので，画面例を用いた説明を省略する．

図 8 は(a)(b)(c)に対応する各画面の後に表示される比較結果の一覧表示画面

図 8. 比較結果表示画面

である．画面は上部，中部，下部に分かれる．まず上部は，比較されている集計群に関する情報であり，調査時期選択画面（図 7）と同様である．画面下部が比較結果を表示する一覧表であり，集計グループと（調査時期ごとの）調査項目のクロス表になっている．クロス表からは，集計グループごとに，各調査時期の値の平均，標準偏差，最小値，最大値，度数がわかる．最下行は調査時期間の値の差を表し，差に関する t 検定により p 値を計算したもの（比較 C の結果）が図の C の位置に表示される(この値は集計グループ 1 のものであるが，集計グループ 2 に関しても同様の位置に対応する p 値が表示されている)．最右列は集計グループ間の差を表し，図の B の位置に同一調査時期の値の差を表す p 値（比較 B の結果）が表示される．（この値は調査時期"1"のものであるが，調査時期"20"に関する値もこのすぐ下に表示されている．）一番右下の A の位置にある値は調査時期間の差に関して，集計グループ間で t 検定を行った p 値（比較 A の結果）である．

画面の中部にある絞り込みインタフェースは，この比較データをさらに絞り込んで分析するためのものである．例えば，男性のみに限った結果や，年齢を 45 歳以上に限った結果など，さらに詳細なデータの傾向を調べるために使用できる．

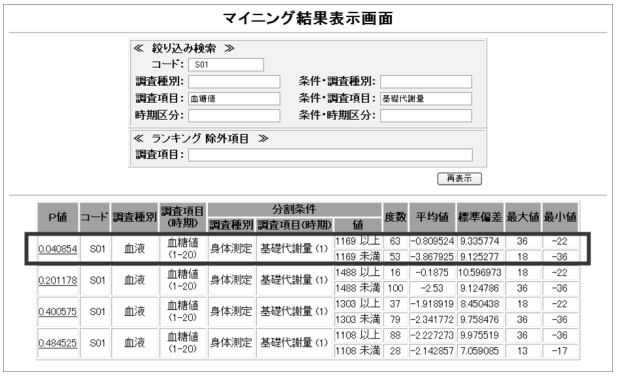

図 9. マイニング画面(比較A)

図 10. マイニング画面(比較 C)

3.4 マイニングインタフェース

本システムでは, 手動の比較インタフェースだけでなく, 効果傾向の異なる興味深い分割条件を自動的にリストアップするデータマイニングインタフェースも備えており, 効率的に効果傾向が把握できる. 比較 A, B, C のそれぞれにマイニングインタフェースが用意されており, 各比較で p 値の低いものから比較条件を表示できる. データ入力時に 2.4 節に示した方法で条件を網羅的に組み合わせて比較を行い, その結果がデータベースに格納されるので, このインタフェースはそのデータを指定条件に応じて表示する機能を持つ.

マイニング結果を p 値の低い順に表示するインタフェースを図 9(比較 A), 図 10(比較 C)に示す. それぞれ, 画面上部が絞り込み検索インタフェース, 下部がリストアップされた結果の一覧表である. 一覧表には, 教室参加者の分割条件（比較 C では「グループ条件」と表記）に加え, 比較した調査項目とその種別, 二つの調査時期（時期 1 と 20 の比較は"1-20"のように表記), そして比較結果を表す統計値が含まれている. p 値の部分はハイパーリンクになっており, 図 8 の比較結果一覧表で詳細な結果を確認できるように配慮されている. 絞り込み検索インタフェースは, 特定の教室や調査項目, 分割条件に関する結果だけを表示したい時に使用する. 図 9,10 はいずれも, 教室コード「S01」の教室に対して,「基礎代謝量」を用いて参加者の分割を行い,「血糖値」の比較結果に差が出たものに絞って表示した例である.

この分析例では, 比較 A の太枠で示した結果を見ると, 教室開始時の基礎代謝量の閾値 1169 で参加者を分割すると, 教室開始時(時期"1")から 2 か月後(時期"20")までの血糖値の変化量には有意な差が見られることがわかる（p 値は約 0.04 であり, 有意水準 5%において有意). この差がどのような差なのかを調べるために比較 C の太枠部分を見ると, 基礎代謝量 1169 未満のグループでは p 値が約 0.0016 であり血糖値の降下に非常に有意な差があるのに対して, 1169 以上のグループでは p 値が約 0.24 と有意な差は見られない. 以上から, この教室の血糖値に対する効果は基礎代謝量が比較的低い参加者のみに現れているという効果傾向が読み取れる.

4. 評価と考察

4.1 評価

本システムのマイニング機能を実際のデータに適用することで, 本システムの効果を評価した. 評価データとして 10 件の教室データをデータベースに登録した. 教室によって実施方法に差があるものの, 1 教室あたり 40～60 名程度の参加者が, 約 3～6 ヶ月の期間参加し

表 1. 比較 A による有意な比較結果数

教室	有意水準	有意な比較結果数	時期の重複を除いた数	時期、分割条件の重複を除いた数
A	1%	47	47	30
	5%	209	209	127
B	1%	146	145	116
	5%	777	762	553
C	1%	142	129	110
	5%	862	759	504
D	1%	5	5	5
	5%	50	50	48
E	1%	448	358	249
	5%	1873	1277	670
F	1%	74	60	46
	5%	643	438	309
G	1%	500	395	306
	5%	2533	1750	1145
H	1%	524	524	324
	5%	1678	1678	1049
I	1%	180	180	97
	5%	498	498	288
J	1%	244	244	187
	5%	926	926	612

た結果である．調査・測定項目については多少のばらつきが見られるが, いずれも身体診断の回数は概ね 3～4 回，アンケート，血液・尿検査は概ね 2 回程度の実施であった．傾向の発見手順として，まずはマイニングインタフェース（比較 A）を用いて群間有意差のある結果を発見し，その後比較 B, C や手動比較画面により詳細な傾向を確認するという方法を用いた．この方法により，どの程度効率的に効果傾向の抽出が可能かを評価した．

まず，マイニングインタフェース(比較 A) により，詳細確認の候補となる，有意な比較結果を列挙した. 各教室について，有意水準 1%と 5%（t 検定の p 値が 0.01, 0.05 以下）とした場合の有意な比較結果の数を表 1 に示す. 教室によりばらつきがあるが，有意水準 1%とした場合でも，最大で約 500 件の，有意差が認められる比較結果が抽出された. これは, 医療専門家が逐一確認するには数が多い．しかし，一方で調査項目数が合計で 200 以上あることを考えると，2 項目の全組合せ数と比べて非常に少数の比較結果が抽出されており, 効果傾向の絞り込みの役割は果たしていると判断できる．

ところで，これらの比較結果は，時期の重複と分割条件の重複の両方を含んでいる．時期の重複とは，比較時期のみが少し異なり, 分割条件や効果のあった調査項目は同じである比較結果である．例えば図 2 は, 時期 1 と 2 に関して年齢と体重の関係を示す比較結果であるが，もし時期 2（または時期 1）のみが異なる比較結果があったとすると, これらは本質的に同一の効果傾向を表しており，この傾向を重複して列挙していると考えられる. 同様に, 分割条件の重複とは，分割に用いるしきい値のみが少し異なる比較結果である．（例えば，図 2 では年齢 50 歳を分割条件として比較であるが，年齢 60 歳を分割条件とした比較があれば, これを分割条件の重複とする.）これらの重複を除いた数が本質的な 2 項目間の関係の数を示していると考えられる．表 1 には，これらの重複を除いた比較結果数も示されているが, 重複を除いても，有意水準 1%で有意差のある比較結果がまだ最大 300 件程度存在し，利用者が逐一確認するには数が多い. 以上のことから, 一つの教室は本質的に異なる多数の効果傾向を内包していることがわかる．

次に, 抽出された傾向のそれぞれについて, 他の比較インタフェースを用いて詳細に検討したところ, 医療専門家でない我々が判断できる範囲内では, 非自明の興味深い結果と思われるものが複数発見され, 本システムが新たな知識を発見できる可能性を示唆できた．例えば，血液に関連する調査項目として，HDL（善玉）コレステロールと LDL（悪玉）コレステロールの値に対して明らかな効果の違いが見られた. それぞれの効果の違いについて詳しく見たところ, ある教室では HDL コレステロールの値に対して, 睡眠時間が 7 時間以上の参加者には明らかな減少効果が見られ, 睡眠時間が 7 時間未満の参加者に効果が見られない現象であったことがわかった. 一方, また別の教室では LDL コレステロールの値に対して, 睡眠時間が 7 時間未満の参加者には明らかな減少効果が見られ，睡眠時間が 7 時間以上の参加者には効果が見られない現象であったことがわかった. この傾向の信頼性については今

後の検証が必要であるものの,未だにあまり知られていない新たな傾向を発見できる可能性が示唆された.

4.2 考察

比較 A において有意な差が見られる比較結果は予想よりも多く，有意水準1%としても概ね数百件以上の（有意差のある）比較結果が得られた．時期や調査項目の重複もそれほど多くはなく,それぞれが教室実施による互いに異なる効果傾向を表す比較結果であることが示唆された．しかし実際には，ユーザである医療関係者がこれらの比較結果を逐一確認していくことは,作業量の面から現実的ではないと思われる.このため,実用化のためには,この中でも特に重要な傾向を抽出する，或いは，比較結果全体をうまく可視化して全体として活用する方法を考案するなどして,現場で利用するにあたっての作業量をさらに削減することが望まれる.

現場で実用性の高い比較結果のみに,自動的に絞りこむ方法の一つとして,医療分野の専門家が知っている既知の効果傾向を知識として登録しておき,これと合致する比較結果を表示しない方法が考えられる.このような既存知識の活用のためには,今後は本システムの結果を医療専門家に検討してもらい,この中の既知の知識を列挙してもらう作業が必要である.このような手法でどの程度の件数の削減が可能であるかは非常に興味深く，今後に是非実施したい.

ところで,上記のように有意な比較結果数が多いという課題はあるものの,一方では,ある程度手間を惜しまなければ,既知でない興味深い関係も発見できる可能性が示唆されたと言える．一例は4.1 節に示したが，このような興味深い関係は他にも見つかっており,本システムの有用性を示す結果となっている.

作業量の面では,確かに本システムを利用すると,数百の比較結果を確認する場合も発生し,作業量は決して小さくはない．しかし，Excel や既存の統計ソフトを用いて同様の解析を行うことと比べると,作業量は各段に減っていることは間違いない.既存の統計ソフトを用いて今回得られたような知見を得るためには,図 2 に示す比較を網羅的に行う必要があり,（1 つの教室の分析だけでも,分割項目，分割しきい値，比較項目，及び２つの調査時期の全組み合わせを試すことになる）とても手作業で行える量ではない.

さらに異なる観点から本システムを見てみると,現在のシステムは各教室の介入内容をシステム上で扱わない点が特徴的である.教室の効果は単に運動指導,食事指導などの大きな分類のみに影響されるわけではなく,講師の印象や地域性など，幅広い要素に影響される．このため，現在のシステムは，教室内容を知っている現地担当者に利用してもらい,得られた傾向と担当者が把握している介入内容との関係性を現場で判断してもらうような利用方法を想定している．しかし，このためには，教室間の違いを直観的に表現するような何らかの可視化機構やインタフェースが必要であろう.この点も今後に向けた改善の一つの方向性になると考えている.

5. おわりに

本研究では,膨大に蓄積されるであろう生活習慣病予防教室のデータから,有用な効果傾向を読み取るデータマイニングシステムを設計・構築した．実際の教室データを用いて評価を行った結果,有意差のある効果傾向を多数列挙することができ,中でも特に興味深いと思われる効果傾向を複数件発見することができた.

今後の課題は,分析者の作業量をより減少できるような比較結果の提示法を提案すること,医療専門家による比較結果の検討を通じて,よりニーズに合った有用なシステムへと改善すること,などが挙げられる.

謝辞

本システムの構築にご尽力いただき，多大な貢献をいただいた南方邦哉氏に心より感謝いたします.

参考文献

[1] メタボリックシンドローム診断基準検討委員会, “メタボリックシンドロームの定義と診断基準,” 日内会誌, Vol. 94, pp.188-203, 2006.
[2] 厚生労働省, “平成 17 年度国民健康・栄養調査結果,” 2007.
[3] 厚生労働省保険医療局生活習慣病対策室, “平成 14 年糖尿病実態調査報告書,” 2003.
[4] National Diabetes Fact Sheet, Department of Health and Human Services of U. S. A., 2005.
[5] 日立メディコ, “生活習慣病リスクシミュレーション,” http://www.hitachi-medical.co.jp/news/news/news0702121.html（2009 年 3 月 22 日参照）
[6] M. Suka, K. Yoshida, and H. Sugimori, “Parsistent Insomnia is a predictor of hypertension in Japanese male workers,” Journal of Occupational Health, Vol. 45, pp.344-350, 2003.
[7] 竹内裕之, 児玉直樹, 橋口猛志, 林同文, “個人健康管理を目的とした健康データマイニングシステム,” 電子情報通信学会データ工学ワークショップ DEWS2006 1B-i11, 2006.
[8] 竹内裕之, 児玉直樹, 橋口猛志, 林同文, “個人健康管理システムのための自動相関ルール抽出アルゴリズム,” 日本データベース学会 DBSJ Letters, Vol. 5, No. 1, pp.25-28, 2006.
[9] M. Pagano and K. Gauvreau, “生物統計学入門,” pp. 186-205, 2003.

（2008年11月17日受付）
（2009年3月25日採択）