健康データマイニングの評価（II）

―― 長期間の時系列データに基づく事例 ――

竹内裕之・児玉直樹・高橋真悟

（受理日　2014年9月24日，受稿日　2014年12月18日）

Valuation of Healthcare-Data-Mining（II）

―― Examples on the basis of long-term data ――

Hiroshi TAKEUCHI・Naoki KODAMA・Shingo TAKAHASHI

（Received Sept. 24, 2014, Accepted Dec. 18, 2014）

1．はじめに

ウェアラブルセンサーによりリアルタイムで生活環境における個人の生体情報や運動量を計測し、Bluetoothなどの無線技術によりスマートフォンにデータを伝送して個人健康管理を行うシステムの開発が進んでいる[1]。これらの膨大な量のデータ（ビッグデータ）はクラウドに蓄積され、何らかの処理をしてシステムユーザの健康管理に役立つ情報を提供することが期待されている。最近の国際学会の潮流としても、m(mobile)-healthやp(personalized)-healthといった概念が浸透している[2]。我々はいち早く、クラウドでデータ処理を行う自動健康データマイニングをコア技術とした個人健康管理システムを開発してきた[3,4]。このシステムは、携帯端末を通して入力した個人の日常の生活習慣と健康に関するデータをクラウドに蓄積し、生活習慣と健康状態の相関ルール抽出（健康データマイニング）を行い、その結果を個人の携帯端末から参照できるものである。個人の生活習慣や健康に関するデータを日毎の粒度で時系列的に蓄積することを前提としており、この時系列データを、我々が開発した遅延相関分析法と呼ばれる手法により解析することが特徴になっている。前報[5]では、本学の学生を中心とした個人健康管理システムのボランティアユーザが、2012年6月1日から11月30日までの6か月間に日毎の粒度で蓄積した生活習慣と健康に関するデータに基づき、開発した健康データマイニング手法によって得られた、パターンやルールについて評価した。

本研究では、個人健康管理システムの1人のボランティアユーザによる8年半余りに亘る長期間の蓄積データを対象として、健康データマイニング手法の適用について評価した結果を報告する。

2．研究方法

2．1．対象ユーザ

本研究の対象ユーザは、東京在住の男性、デー

タ取得開始時（2004 年 5 月）57 歳、通常の勤務を行っている。やや血圧に問題があり健康管理に関心をもって日々のデータを取得している。このユーザの血圧、脈拍数、体脂肪率、消費エネルギー、摂取エネルギーの 8 年半余りの長期にわたる日毎の粒度の時系列データを対象に、健康データマイニング手法を評価した。

体重、体脂肪率は、タニタの体組成計（Inner Scan：BC-521）を用い、毎朝起床時に計測した。血圧、脈拍は日本精密測器の血圧計（VITAL SCOPE）を用いてやはり起床時に計り、血圧については 3 回計測してその平均値をデータ登録した。生活習慣としての消費エネルギーは、歩行によるものはオムロンの歩数計（Walking style）を携帯して計測し、その他の運動については Mets 値を基に推測した。摂取エネルギーについては毎食事の内容から、インターネット上の関連サイトを参照するなどして推測した。データの登録・蓄積期間は、2004 年の 5 月から 2012 年の 12 月までで、データは原則的にほぼ毎日システムに登録されていた。

2．2．遅延相関分析法に基づく健康データマイニング

我々が開発している健康データマイニングでは、「生活習慣の蓄積が健康状態に変化をもたらし、その影響は時間遅れをもって現れることがある」という極めてシンプルなモデルをベースとしている。図 1 に示すように、健康状態 h と生活習慣 e の個々の時系列データ間には相関が見られなくても、生活習慣データの蓄積やその健康状態への影響の遅延を考慮するとしばしば健康状態の変化との間に相関がみられることがある[6]。すなわち、健康状態の変化

$$\Delta h_{nm} = h_n - h_m \tag{1}$$

と生活習慣データの蓄積

$$e^t{}_{ij} = e_i + e_{i-1} + \cdots + e_j \tag{2}$$

の間に、遅延期間 $s = n - i \geqq 1$ を考慮すると時系列データ間に隠れていた相関をあぶりだすことができる。相関の評価には次式で表される時系列データ間のピアソンの積率相関係数を用いる。

$$r(\Delta h_{nm},\ e^t{}_{ij}) = \frac{Cov(\Delta h_{nm},\ e^t{}_{ij})}{SD(\Delta h_{nm})\, SD(e^t{}_{ij})} \tag{3}$$

ここで、r は相関係数、$SD(\triangle h_{nm})$ は $\triangle h_{\mathrm{nm}}$ の時系列対象区間における標準偏差、$SD(e^t{}_{ij})$ は $e^t{}_{ij}$ の時系列対象区間における標準偏差、$Cov(\triangle h_{nm,}\ e^t{}_{ij})$ は $\triangle h_{nm}$ と $e^t{}_{ij}$ の共分散である。

具体的には、対象とする健康状態 h と各種生活習慣 e の時系列データについて、n-m，i-j，s をパラメータとして式（3）のピアソンの積率相関係数を評価し、相関係数の絶対値が最大となる（n-m，i-j，s）のセット（$(n\text{-}m)_{\max}$，$(i\text{-}j)_{\max}$，$s_{\max}$）を見出す。そして、相関係数の絶対値がある閾値より大きい場合に、その生活習慣の蓄積を対象とする健康状態の変化に対する説明変数として採用する。例えば、$(i\text{-}j)_{\max} = 2$，$s_{\max} = 2$ で、相関係数が閾値を超えていれば、

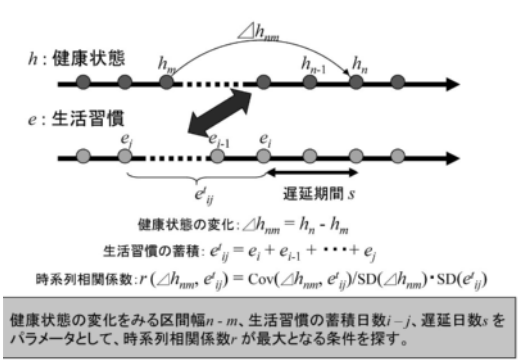

図 1　遅延相関分析法

$$e_i + e_{i-1} + e_{i-2} \quad (i = n-2) \qquad (4)$$

すなわち、「2 日前から 3 日間の生活習慣 e の蓄積」を説明変数のひとつとして採用する。

次に、目的（ターゲット）変数である健康状態に関しては、その時系列データが数値の場合には「高い」「中間」「低い」の 3 つのシンボル値を持つ変数に変換する。このとき、各シンボル値に属する数値データ数がほぼ同数になるように境界値を設定する。そして、説明変数と目的変数の間のルール生成には ITRULE アルゴリズム[7] を用いたアソシエーションルール解析もしくは決定木による解析を行う。両者はほぼ同等なルールを抽出する。我々が開発したクラウド型個人健康管理システムにおける自動健康データマイニングでは前者の手法を採用しているが、本研究では決定木により手動でルール解析を行った。なお、解析に用いたツールは IBM 社の Clementine である。

3．解析結果

3．1．データの季節変動の補正

本研究で対象としたボランティアユーザの長期データに関しては、すでに別報[8] で一部解析を行っており、体脂肪率、血圧、脈拍の時系列データに明瞭な周期的季節変動が観測されている。その約 8 年間に亘るデータの月毎平均をみると、変動幅はほぼ±5 %となっている[8]。従って、長期間のデータを基に生活習慣と健康状態間の相関ルール解析を行うにあたり、季節変動をバイアスとみて補正する必要があると考えた。月毎平均の季節変動幅の実測値に基づいて、表 1 に示したように各月のデータを月毎に異なる率で補正した。

表 1　時系列データの月毎の季節変動補正率

月	体脂肪率の補正率	血圧の補正率
1	0.95	0.97
2	0.96	0.95
3	0.98	0.97
4	1	0.99
5	1.02	1
6	1.04	1.01
7	1.05	1.03
8	1.04	1.05
9	1.02	1.03
10	1	1.01
11	0.98	1
12	0.96	0.99

3．1．1．体脂肪率データの補正

図 2 は、2004 年 6 月 1 日から 2012 年 12 月 31 日の 8 年 7 か月に亘る体脂肪率の補正前と補正後の日毎の時系列データである。補正前には、日毎粒度のデータでみても、夏に低く冬に高いという明瞭な周期的季節変動があることが判る。データを月毎に異なる率で補正した後には、季節変動のバイアスがかなり除かれていることが判る。補正の効果は次の図 3 のヒストグラムにおいて更に明瞭に現れている。補正前は、非対称であった分布の形状が補正後はかなり対称な正規分布に近づいている。

3．1．2．血圧データの補正

最大血圧（収縮期血圧）と最小血圧（拡張期血圧）の 2004 年 6 月 1 日から 2012 年 12 月 31 日に亘る補正前と補正後の日毎時系列データを図 4 および図 5 に示す。どちらも補正前は夏に低く冬に高いという季節変動が明らかに重なっているが、補正後は季節変動のバイアスが抑えられていることが判る。補正の効果は図 6、図 7 のヒストグラムに更に顕著に現れている。季節変動補正により、最大血圧、最小血圧とも形状から判断して正規分布に近づいている。特に、

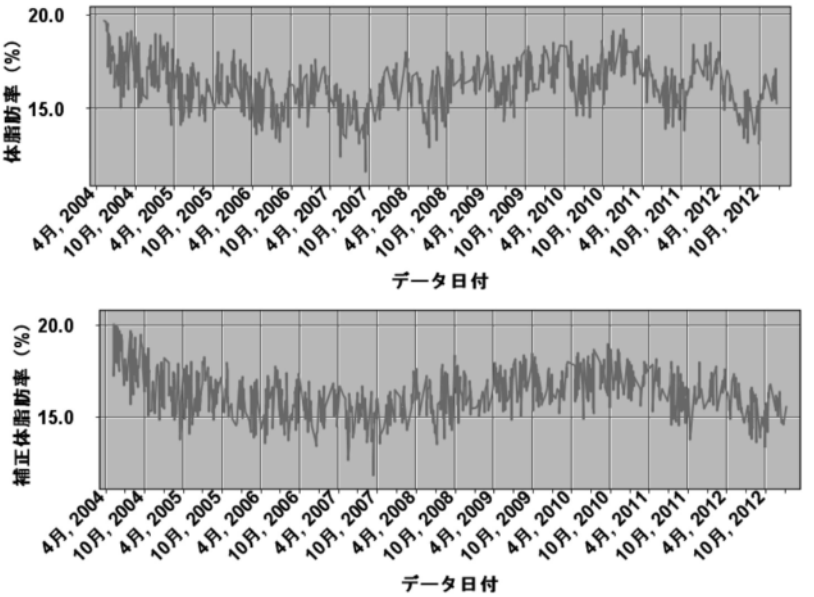

図 2　体脂肪率と補正体脂肪率の時系列変化

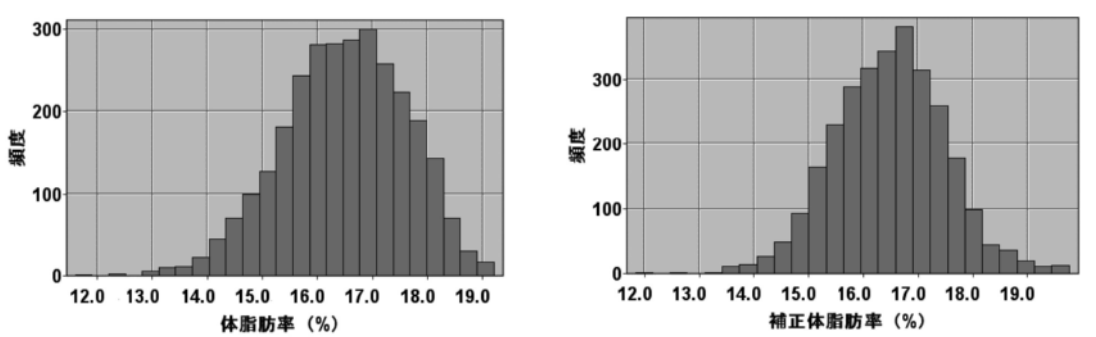

図 3　体脂肪率と補正体脂肪率のヒストグラム
（データ取得期間：2004年 6 月 1 日～2012年12月31日）

最小血圧に関して、補正の効果は著しい。

3．2．遅延相関分析

3．2．1．補正の影響

季節変動補正の前後の時系列データを基にして、それぞれ遅延相関分析を行った結果、季節変動補正は分析の結果には殆ど影響しなかった。例えば、2004 年 6 月 1 日から 2005 年 5 月 31 日までの 1 年間の時系列データを基に、総消費カロリー（歩数消費カロリーとその他運動による消費カロリーの総和）と体脂肪率変化の間の遅延相関分析を行った結果、補正前後双方において 2 日間の総消費カロリーが 2 日遅れで、2 日前からの体脂肪率変化に与える影響が最大になるという結果であった。それぞれの散布図を図 8 に示す。散布図の様相は補正前後で殆ど変らず、評価した相関係数にも有意差はなかった。

この原因は、遅延相関分析において、目的変数である健康データ（この場合には体脂肪率）に関しては常にある期間の変化（差分）が分析

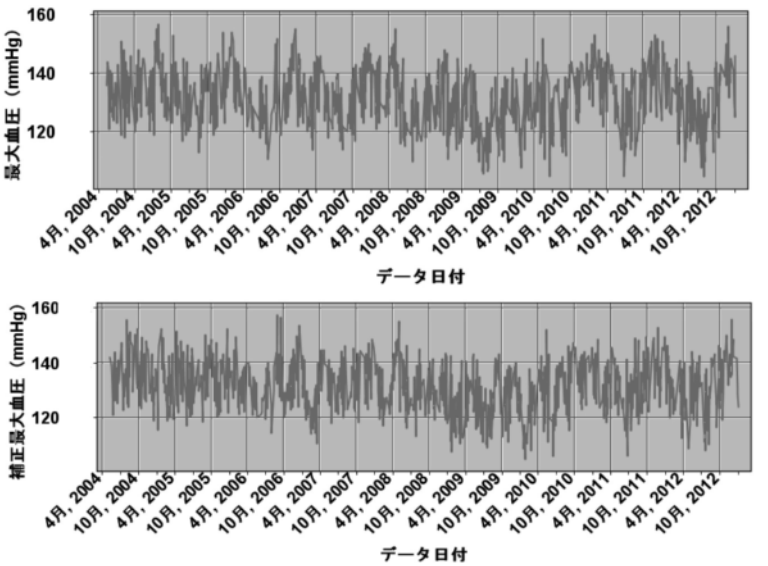

図 4　最大血圧と補正最大血圧の時系列変化

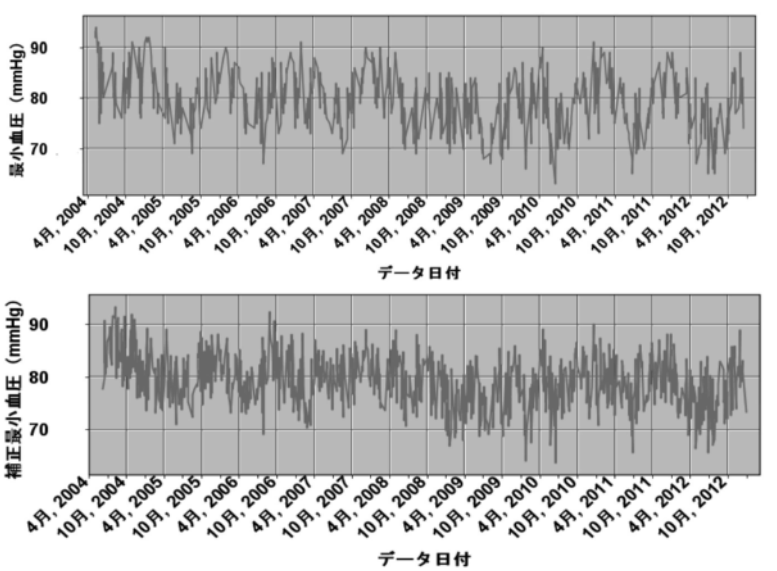

図 5　最小血圧と補正最小血圧の時系列変化

対象であることによる（式 (1)）。すなわち、相関を評価するときに季節変動分は常にほぼ相殺されることになる。特に、本研究の遅延相関分析においては、健康データ変化をみる期間幅（$n-m$）の最大を 10 日間としているので、季節変動の周期に比べて充分短いことが大きな原因になっていると考えられる。

3．2．2．長期間データの分析

遅延相関分析においては、季節変動によるバイアスの影響は殆どないことが判ったので、補正前の生の長期間データを対象に分析を行っ

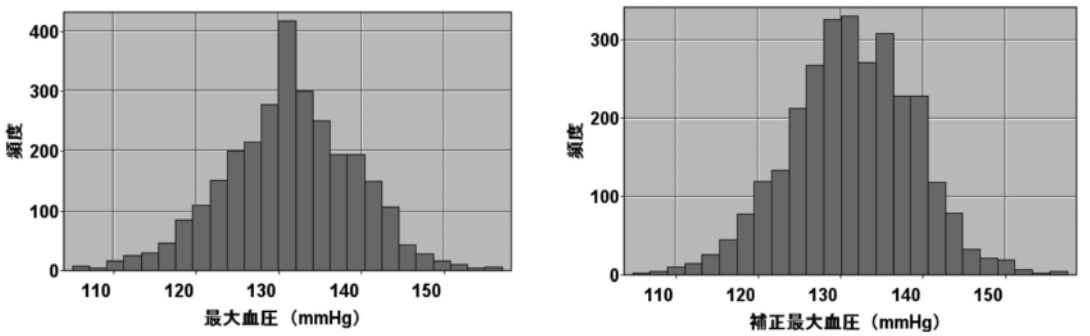

図 6　最大血圧と補正最大血圧のヒストグラム
（データ取得期間：2004年 6 月 1 日～2012年12月31日）

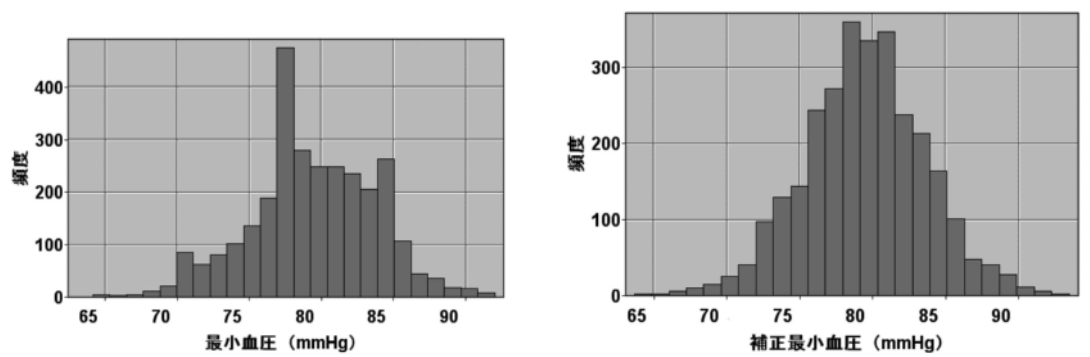

図 7　最小血圧と補正最小血圧のヒストグラム
（データ取得期間：2004年 6 月 1 日～2012年12月31日）

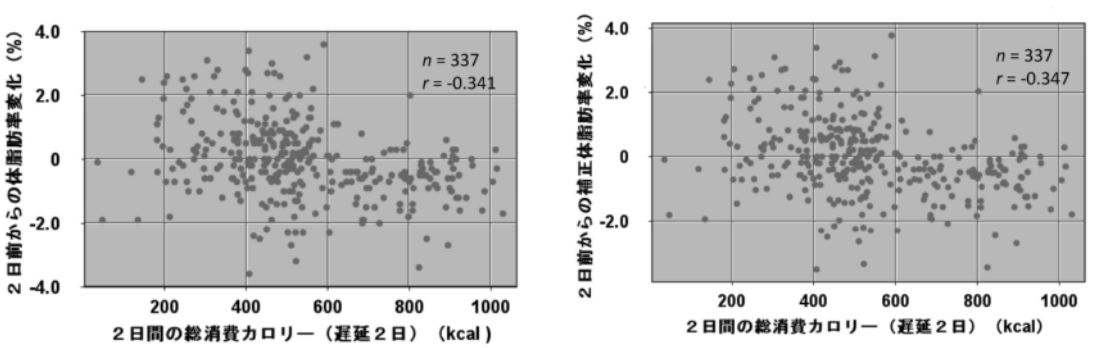

図 8　（補正）体脂肪率変化と総消費カロリーの散布図
（データ取得期間：2004年 6 月 1 日～2005年 5 月31日）

た。図 9 は 2004 年 6 月 1 日から 2005 年 5 月 31 日までの 1 年間のデータを基に、総消費カロリーと体脂肪率の相関分析を行った結果（散布図）であるが、前項で述べたように、2 日間の総消費カロリーと 2 日前からの体脂肪率変化の間に遅延 2 日の場合に有意な負の相関がみられる。しかし、遅延日数を前後に 1 日変えてみると、相関を全く示さなくなることが判る。

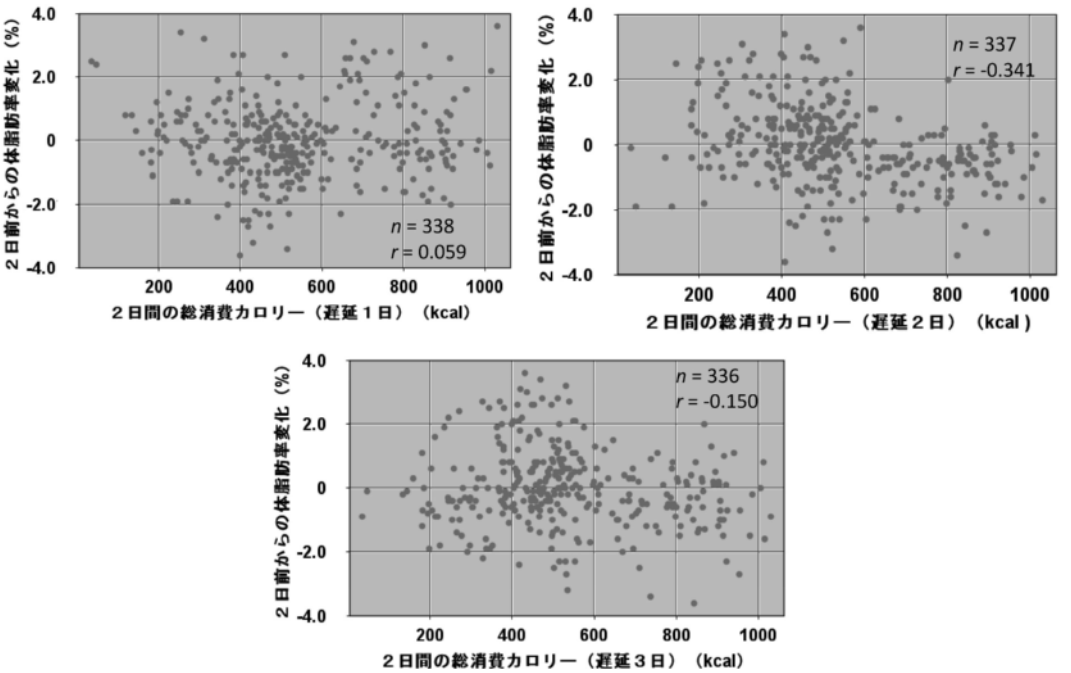

図９　体脂肪率変化と総消費カロリーの散布図の遅延日数による変化
（データ取得期間：2004年６月１日～2005年５月31日）

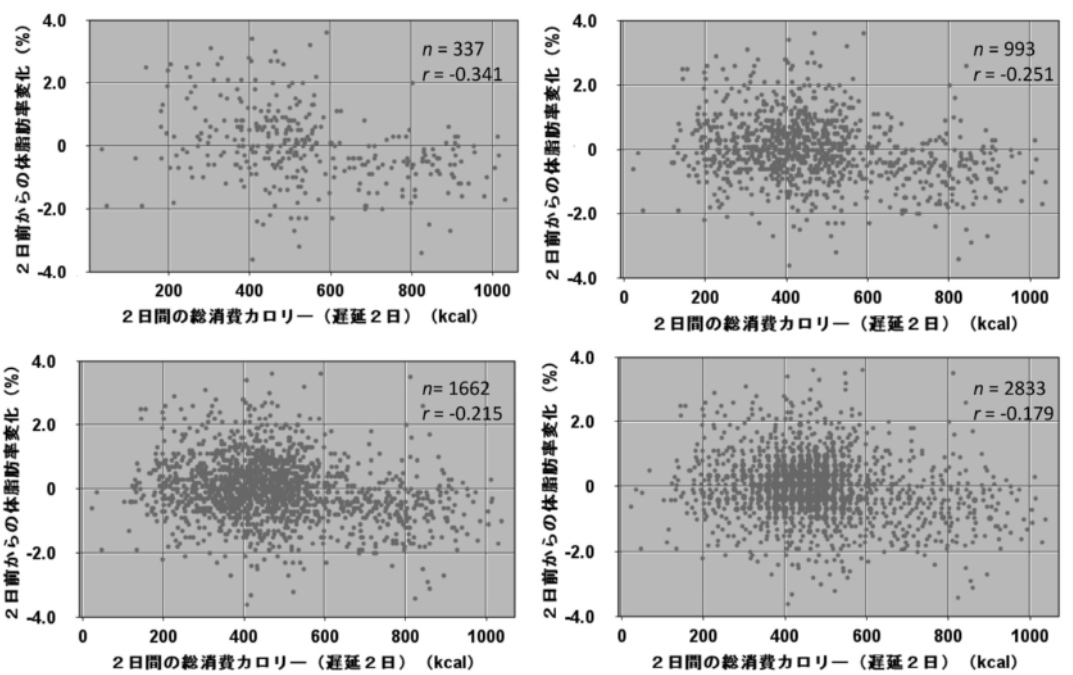

図10　体脂肪率変化と総消費カロリーの散布図のデータ蓄積期間による変化

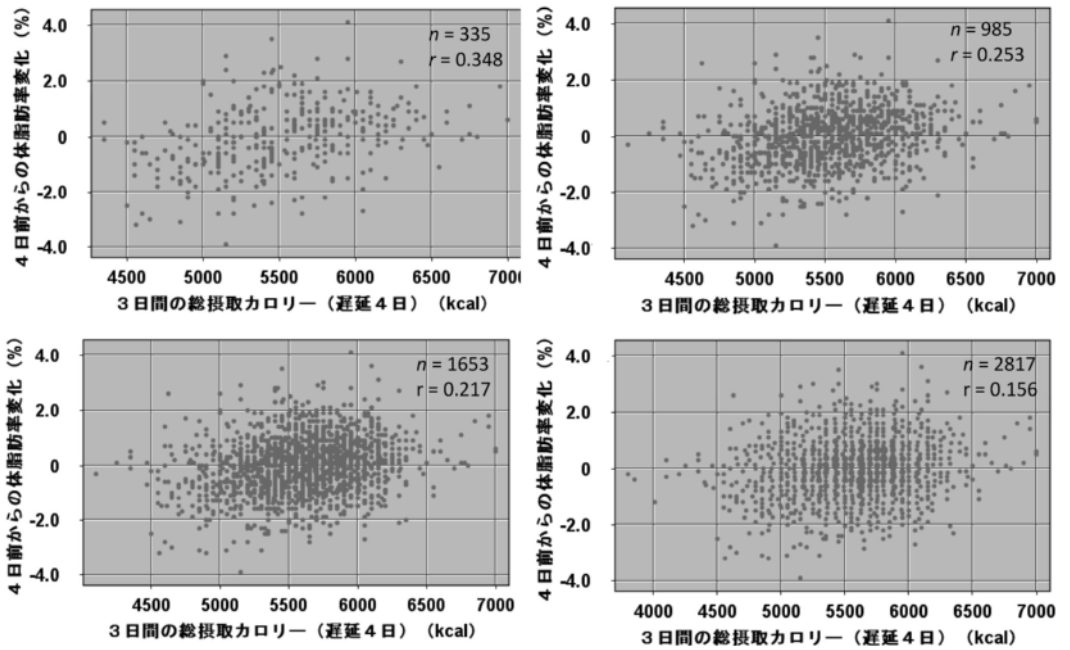

図11　体脂肪率変化と総摂取カロリーの散布図のデータ蓄積期間による変化

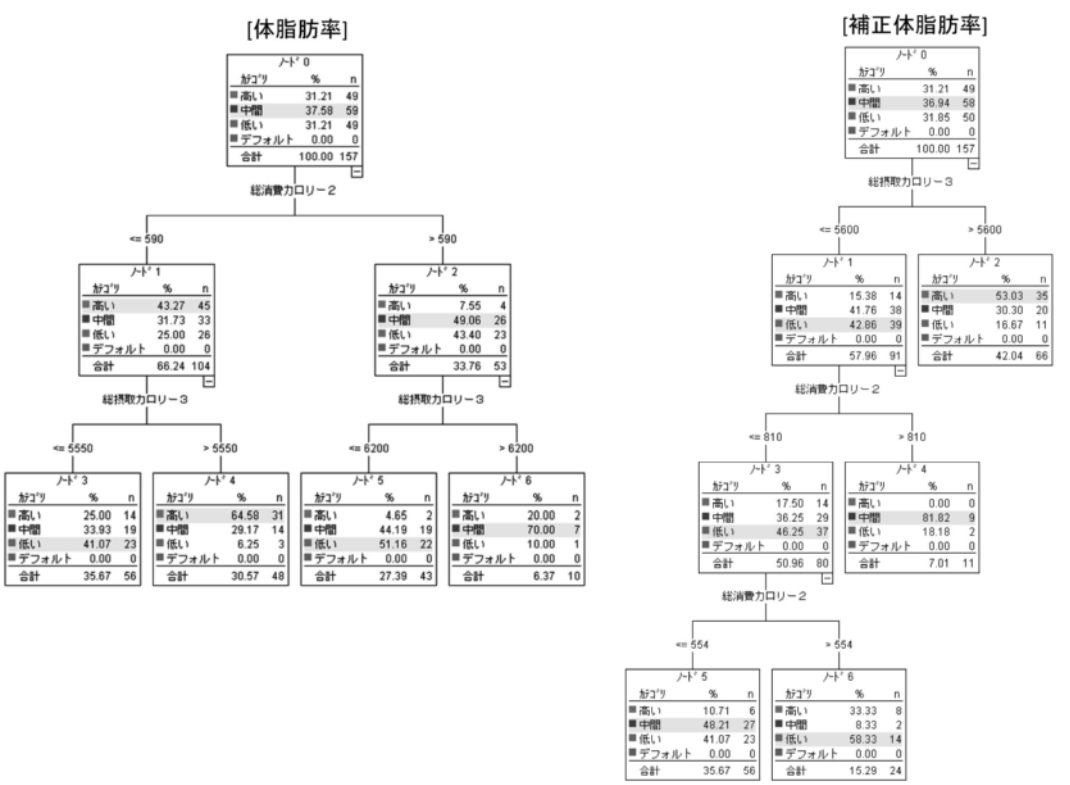

図12　（補正）体脂肪率を目的変数とした決定木
（データ取得期間：2004年６月１日～2004年11月30日）

[目的変数:体脂肪率]

総消費カロリー2 <= 590 [モード: 高い]
　総摂取カロリー3 <= 5550 [モード: 低い] => 低い
　総摂取カロリー3 > 5550 [モード: 高い] => 高い
総消費カロリー2 > 590 [モード: 中間]
　総摂取カロリー3 <= 6200 [モード: 低い] => 低い
　総摂取カロリー3 > 6200 [モード: 中間] => 中間

[目的変数:補正体脂肪率]

総摂取カロリー3 <= 5600 [モード: 低い]
　総消費カロリー2 <= 810 [モード: 低い]
　　総消費カロリー2 <= 554 [モード: 中間] => 中間
　　総消費カロリー2 > 554 [モード: 低い] => 低い
　総消費カロリー2 > 810 [モード: 中間] => 中間
総摂取カロリー3 > 5600 [モード: 高い] => 高い

図13　（補正）体脂肪率を目的変数とした決定木から抽出されたルール（データ取得期間：2004年 6 月 1 日～2004年11月30日）

[目的変数:体脂肪率]

総摂取カロリー3 <= 5550 [モード: 低い] => 低い
総摂取カロリー3 > 5550 [モード: 高い]
　総消費カロリー2 <= 608 [モード: 高い] => 高い
　総消費カロリー2 > 608 [モード: 中間] => 中間

[目的変数:補正体脂肪率]

総摂取カロリー3 <= 5550 [モード: 低い] => 低い
総摂取カロリー3 > 5550 [モード: 高い]
　総消費カロリー2 <= 658 [モード: 高い] => 高い
　総消費カロリー2 > 658 [モード: 中間] => 中間

図15　（補正）体脂肪率を目的変数とした決定木から抽出されたルール（データ取得期間：2004年 6 月 1 日～2004年 8 月31日）

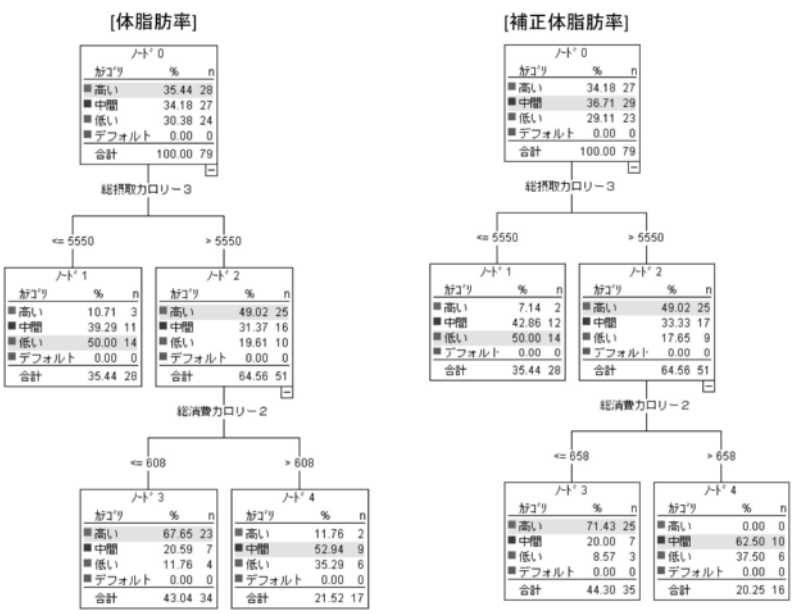

図14　（補正）体脂肪率を目的変数とした決定木
（データ取得期間：2004年 6 月 1 日～2004年 8 月31日）

また、遅延 2 日の散布図を観察すると、2 日間の総消費カロリーが 600kcal 前後を境に様相が変わっており、それより少ないところではプラス変化とマイナス変化の確率はほぼ同等であるが、多いところではマイナス変化の確率が圧倒的に優位になるという特徴がみられる。同様な解析を 2004 年 6 月 1 日から 2007 年 5 月 31 日までの 3 年間、2009 年 5 月 31 日までの 5 年間、2012 年 12 月 31 日までの 8 年 7 か月間におけるデータについて行った結果を図 10 に示す。1 年間のデータに基づく散布図に見られた特徴が加齢を重ねても受け継がれていくことが判る。

同様な分析を、総摂取カロリー（1 日の合計摂取カロリー）と体脂肪率の間で行った結果を図 11 に示す。2004 年 6 月 1 日から 2005 年 5 月 31 日の 1 年間のデータでは、3 日間の総摂取カロ

リーと 4 日前からの体脂肪率変化の間に遅延 4 日の場合に有意な正の相関がみられ、その特徴はやはり加齢とともに受け継がれていくようにみえる。

3．3．ルール生成

遅延相関分析においては季節変動の影響は殆どないことが判ったので、健康データマイニングの前段処理である説明変数の選択は季節を跨る長期間のデータが対象でも従来通り実行できる。しかし、後段処理のルールマイニングにおいては、目的（ターゲット）変数である健康状態を、データが数値の場合には「高い」「中間」「低い」の 3 つのシンボル値を持つ変数に変換し、それぞれの値に属するデータ数がほぼ等しくなるように境界値を設定するので、データ値の季節変動補正の影響を直接受ける。

決定木を用いてルールを生成した実例を以下に示す。図 12 は、すでに説明した 2 日間の総消費カロリー（遅延 2 日）と 3 日間の総摂取カロリー（遅延 4 日）を説明変数として採用し、2004 年 6 月 1 日から 2004 年 11 月 30 日までの半年間の体脂肪率の補正前および補正後のデータをターゲットとしてそれぞれ作成した決定木である。そして、図 13 はこれらの決定木から導かれたルールである。補正前後の体脂肪率のデータは、同じデータ数割合で「高い」「中間」「低い」に分類されているが、得られた決定木とルールは補正前後で異なっている。抽出されたルールを良く吟味すると、補正前も補正後もほぼ同じ傾向が読み取れるのであるが、決定木のルートノードから最初の分割に用いる説明変数がそもそも異なっている。

試行中のクラウド型個人健康管理システムでは、このような季節変動の影響を意識し、1 シーズン（3 か月間）のデータを基に自動健康データマイニングを実行しており、前報[5]では 3 か月間のデータを基に行った結果を評価した。そこで、本研究でも 2004 年 6 月 1 日から 2004 年 8 月 31 日までの 3 か月間のデータを基に、補正前後で決定木によるルール生成を実行してみた。図 14、15 はその結果である。補正前後の体脂肪率データは同じデータ数割合で「高い」「中間」「低い」に分類されており、総消費カロリーの値でノード 2（図 14）を分割するときの境界値が若干異なるものの、全く同じ形の決定木が得られている。そして、その帰結として同じようなルールが得られている。

4．考　察

4．1．季節変動の補正とデータの正規化

1 人のボランティアユーザの 8 年 7 か月に亘る日毎粒度の体脂肪率、血圧、脈拍などのデータが季節によって周期的な変動を示し、生活習慣とこれら健康状態の相関ルール解析に影響を与えるであろうことは別報[8]で指摘した。実際に 3．1．で示したように、8 年 7 か月に亘る 3000 近い体脂肪率および血圧のデータの分布には明らかな偏りがあるが、季節変動補正をかけることにより、正規分布に近いものとなった。このことは、季節を跨る長期間のデータをもとに生活習慣とこれら健康状態の相関ルール解析を行う場合には、季節変動の補正を行うべきであることを示唆している。

同時に、季節変動というバイアスを除くと、健康である限りヒトのバイタルサインの時系列のばらつきは平均値を中心に正規分布に近いという結果を示している。季節の変化に応答しながらも生体は恒常性（ホメオスタシス）を保っ

ていることの表れではないかと考えられる。

4．2．長期間データの遅延相関分析

体脂肪率や血圧の季節変動は、本研究で行っている遅延相関分析には影響を与えないことがわかった。

この結果を背景に、ボランティアユーザの長期に亘る生データを解析した結果、消費カロリーや摂取カロリーなど生活習慣の蓄積と体脂肪率変化の相関には、有意な相関を示す遅延日数や蓄積期間に個人の特徴が現れ、そのパターンは加齢を重ねても受け継がれていくという傾向がみられた。これは、有意な相関を示す遅延日数や蓄積期間は遺伝的な特質であり、季節や加齢に影響されないことを示していると考えられる[9]。健康データマイニングの観点からは、加齢とともにルールそのものは変化していっても、ルールマイニングに用いる説明変数は個人毎に長期に亘り同じであることを示している。

4．3．季節変動補正とルールの生成

健康データマイニングでは、目的変数（健康状態）が数値データの場合に、属するデータ数がほぼ同じになるように、「高い」「中間」「低い」という 3 つのシンボル値を持つ変数に変換してルールを生成するので、データの季節変動補正は当然ルール生成に影響を与える。3．3．ではその例を示したが、補正によりデータは正規分布に近づくのであるから、原則的に補正後のデータを基にルール生成すべきである。しかし、本研究で示したように個人データの季節変動の詳細は長期間のデータ蓄積の結果判るものであり事前には判らない。つまり、データ補正は研究としては意味があるが、試行中のクラウド型個人健康管理システムに実装することは困難である。幸い、図 14、15 に示すように、1 シーズン（3 か月間）のデータを解析対象にするのであれば、当然のことではあるが季節変動補正の有無はあまりルール生成に影響しないので、現在試行しているように 3 か月間の補正無しの生データをもとに健康データマイニングを行うのは現実的な手法のひとつである。

5．まとめ

試行中のクラウド型個人健康管理システムの 1 人のボランティアユーザの 8 年 7 か月に亘る日毎粒度の長期間データに基づき、開発中の健康データマイニング技術を評価した。その結果を以下に要約する。

(1) 体脂肪率、血圧などのデータは明らかな周期的季節変動を示し、その変動幅を基にデータ補正を施すことにより、データのばらつきは正規分布に近いものになった。

(2) 健康データマイニングの核となる遅延相関分析においては、健康データの季節変動はほぼ相殺されるため、季節を跨る長期間のデータを対象にしてもその妥当性は保証される。

(3) ボランティアユーザの最初の 1 年間のデータを対象に遅延相関分析により得られた、総消費カロリーおよび総摂取カロリーと体脂肪率変化の間に最大の相関をもたらす遅延日数などの特徴は、その後 8 年余りの期間において加齢とともに受け継がれていく。

(4) 季節を跨る期間のデータを対象にルール生成を行う場合には、季節変動補正後のデータを用いるべきであるが、データ補正処理をクラウド型個人健康管理システムに実装

するのは困難であり、現行システムで実施しているように1シーズン（3か月間）のデータに基づき、ルール生成をおこなうのが現実的である。

謝辞

本研究は文部科学省科研費（課題番号：26350868）の助成を受けている。また、日本データベース学会と日立製作所による日立HiRDBアカデミック制度の適用を受けている。

参考文献

1）E. Kantoch, P. Augustyniak, M. Markiewicz, and D. Prusak : Monitoring activities of daily living based on a wearable wireless body sensor network, Proc. 36th Annual International Conference of the IEEE EMBS (2014) 586-589.

2）B. C. Zapata, A. H. Ninirola, J. L. Fernandez-Aleman, and A. Toval : Assessing the Privacy Policies in Mobile Personal Health Records, Proc. 36th Annual International Conference of the IEEE EMBS (2014) 4956-4959.

3）H. Takeuchi, T. Hashiguchi, and T. Shintani : Personal Dynamic Healthcare System Utilizing Mobile Phone and Web Technologies, Proc. 2nd Int' l Conf. Advances in Biomedical Signal and Information Processing (2004) 304-307.

4）竹内裕之、児玉直樹、橋口猛志、林　同文：インターネット上で動く自動健康データマイニングシステム、高崎健康福祉大学紀要　第5号（2006）1-11.

5）竹内裕之、児玉直樹：健康データマイニングの評価（Ⅰ）―6ヵ月間の時系列データに基づく事例―、高崎健康福祉大学紀要　第13号（2014）1-8.

6）竹内裕之、児玉直樹：生活習慣と健康状態に関する時系列データ解析手法の開発、第19回データ工学ワークショップDEWS 2008 論文集（2008）E1-5.

7）P. Smyth and R. M. Goodman : An Information Theoretical Approach to Rule Induction from Databases, IEEE Trans. Knowledge and Data Engineering, vol.4, no.4 (1992) 301-316.

8）竹内裕之、黛　勇気、児玉直樹：健康と生活習慣に関わる時系列データ解析に基づくp-healthの1例、高崎健康福祉大学紀要　第12号（2013）11-19.

9）H. Takeuchi, Y. Mayuzumi, and N. Kodama : Parameters Characterizing Nature of Personal Health in the Correlation between Energy Expenditure/Supply and Body- Fat, Proc. 34th Annual International Conference of the IEEE EMBS (2012) 2140-2143.