2000 1141

厚生科学研究研究費補助金

医療技術評価総合研究事業

標準データ項目セットを用いた知的データベースによる診療根拠の動的生成に関する研究

平成１２年度　総括研究報告書

主任研究者　山本　隆一

平成１３（２００１）年４月

目　　次

厚生科学研究費補助金(医療技術評価総合研究事業)

総括研究報告書

標準データ項目セットを用いた知的データベースによる診療根拠の動的生成に関する研究

総括研究者　山本　隆一　大阪医科大学病院医療情報部助教授

研究要旨

ＥＢＭが重要であることは論をまたないが、一般的な文献を基礎とする手法では一過性で流行性の疾患や薬剤の副作用の速やかな発見などには適用できない。そこで診療情報の電子化が行われることを前提に、それらの情報をプライバシー保護を確保して上で収集し、データマイニングの手法を用いてダイナミックに知識を抽出し診療根拠とすることを目標とし、そのための基礎研究をおこなった。平成１２年度はデータマイニングの手法として相関ルールとそのための前処理に関して検討し、データ項目セットの適応性、およびプライバシー保護と安全性確保のためのデータの無名化の検討と、プライバシー・センシティブな情報の扱いについて研究をおこない一定の成果を得た。

分担研究者：

大江和彦

東京大学医学部附属病院中央医療情報部教授

坂本憲広

九州大学医学部附属病院　講師

A. 研究目的

本研究の目的は平成１１年度の厚生省によるデータ項目セット開発事業の成果を利用し、この項目セットに準拠して収集された診療データから動的に診療根拠を抽出する方法を研究開発することにある。

証拠に基づく診療（EBM）が重要であることは論を待たないが、一般的なEBMのように文献的な証拠に基づく場合、インフルエンザに対する抗ウイルス剤の効果のような流行性で急性の疾患への対処や、薬品の副作用などの迅速な対応を必要とする場合などでは十分な効果が期待できないことがある。例えば今年のインフルエンザにアマンタジンが効果を示すかどうかといった

場合、文献的な根拠を待つことができないために、診療現場からの経験が厳密な検証なく、また統計的な処理がほとんど行われずに流布する形でわずかに現場医療に活かされている状況にある。薬品の副作用も相当な例数の蓄積と回顧的な解析が必要であるが、副作用を疑う医療現場からの報告に依存しており、疑うことが難しい状況では調査そのものも遅れる可能性がある。そしてこれらの場合、現場の印象がトリガーになる。熟練した医療従事者の印象は実際には複雑な知識背景のもとに下される判断で、高く評価する必要があるが、客観性は不十分といわざるを得ない。

平成１１年度に開発されたデータ項目セットは電子化診療情報を共通の標識で整理することで、異なる医療機関の診療情報を統一的に扱うことを可能とするもので、これを活用することにより、広範囲から診療情報をリアルタイムに収集することが可能になる。

一方で情報工学の分野では知的データベースやデータ・マイニングと呼ばれる手法の研究が活発に行われている。これは網羅的に集められたデータの集合から意味のある関係を自動的に抽出する手法であり、人の印象に頼らないデータ解析を行うことができる。着眼点を指定しなくても動的に特異な関係を抽出できるために人が気付き難い関係や、気付くのに時間がかかる関係を早期に抽出するのに極めて有用な方法と考えられる。もちろん従来の回顧的な方法にくらべて若干の精度低下は予想されるし、背景となるべき医学理論までを類推することはできない。あくまでもヒントを与える方法と考えることができ、回顧的な研究方法を併用する必要がある場合が多い。しかし少なくとも着眼点を得るまでの時間は大幅に短縮され、医療現場へのフィードバックもそれだけ有効になることが期待できる。

しかしデータ項目セットは開発されたばかりで、実装応用例は少なく、このような目的のデータ収集に十分な能力があるか検証する必要がある。またデータマイニングは情報工学の分野で活発に研究されており、いくつかの分野では実用的な応用例も存在するものの、医療分野での定まった適用方法はない。データマイニングにはさまざまな手法があり、またその前提となるデータの前処理にも定まった方法はない。医療で実用的に用いるためには実証的な研究が必要である。また動的に診療情報を収集する以上はプライバシー保護はきわめて重要な課題となる。そこで本研究ではこのような用途に関してのデータ項目セットの検討を分担研究者の大江を中心に行い、プライバシー・センシティブな

情報項目の扱いを分担研究者の坂本を中心に行い、データマイニングおよび前処理の手法の研究とデータ収集の際のプライバシー保護を含めた安全性確保の研究、およびこれらの総括を山本が行うこととした。

データマイニングおよび前処理については本年度は、既に収集された静的な医療データベースから、診療根拠につながるルールを発見する手法について研究を行った。医療データベースに対し既存のデータマイニング手法を実際に適用することにより、現状で適用可能なデータマイニング手法と医療データベースへ適用する際の問題点を明らかにし、次年度以降の研究の基礎とした。

安全性確保の研究については実際の患者基本情報データベースから実験用に変換したデータを用い、データの無名化の定量化を試みることとした。

データ項目セットおよびプライバシー・センシティブな項目の扱いについては分担研究報告書を参照されたい。

B. 研究方法

１．無名性の定量化について

ａ　大阪医科大学付属病院に過去５年間で利用された患者情報３２万件を用い、厚生労働省の補助で作成された「電子化された診療情報交換のためのデータ項目セット(以下J-MIXと呼ぶ)」にしたがって項目整理を行った。

ｂ　上記のデータベースを用い、単一項目８１４項目について最小特定人数を計算した。また郵便番号は上３桁のみ、住所は市や町名レベルでも計算した。また年齢は階層でも計算し、生年月日は月単位、年単位でも計算した。

ｃ　２の単一項目のうち、よく使われると思われる組み合わせ８０組について最小特定人数を計算した。

なお、本研究はデータの無名性を研究するものであり、研究のためのデータは実データを使用した。しかし、特定の程度だけを指標として用いるために、何に特定されるかは問題ではない。そこで、氏名などの特定性の強い項目ははじめから除外して扱った。

２．データマイニングおよび前処理について

ａ　相関ルール発見

大量に蓄積された診療データからルールを発見する手法として、データ中の属性間の相関関係を抽出する相関ルール発見手法を適用する。データの属性集合をI、データベースをDとすると相関ルールは$X \Rightarrow Y$で表現され、$X, Y \subset I$、$X \cap Y$＝空集合である。相関ルールはルールが持つ支持度(support)と確信度(confidence)の2つの

値を使ってその有意性を示す。相関ルールX ⇒ Yの支持度support(X ⇒ Y)はD全体に対しXとYをともに含む事例の割合support(X ∪ Y)と定義される。確信度confidence(X ⇒ Y)はD中でXを含む事例のうち、XとYを共に含む事例の割合、すなわちsupport(X ∪ Y) / support(X) によって定義される。相関ルール発見とは、ユーザが指定する最小支持示と最小確信度を満足するすべてのルールを抽出する問題となる。相関ルール発見の効率的なアルゴリズムの1つに、IBMアルマデン研究所のAgrawalらが提案したアプリオリアルゴリズムがある。これは、最小支持度の性質を利用し探索空間を枝刈りしながら長さ１のルールから順に条件を満たす相関ルールを抽出する。本研究ではこのアプリオリアルゴリズムを適用し、診療データから相関ルールを発見する。

b　データの前処理

対象となる診療データベースとして、大阪医科大学付属病院で実際に蓄積された糖尿病データベースを用いた。本データベースは、60個の属性から構成されており1251の事例を含んでいる。60個の属性のうち、「年齢」や「Body Mass Index」といった属性値として数値をとる連続属性は31個存在する。一方、「網膜症有無」や「喫煙歴有無」といった、属性値としてあらかじめ定義された離散値をとる離散属性は29個である。それぞれの属性は欠損値を含む。これらのすべての属性を使用しアプリオリアルゴリズムをそのまま適用すると、大量のルールが生成され、ルールの発見にかかる時間も大きくなる。そこで、相関ルール発見手法をより効果的に適用するために、前処理として以下の処理を糖尿病データベースに対して施した。

○知識発見に無関係と思われる属性を排除

「患者ID」や「カルテNo」といった知識発見に本質的に関係無い属性を排除した。

○欠損値を多く含む属性を排除

属性の中には、1251事例のうち1000個以上もの事例について、その属性値が欠損しているものが存在した。そのように、欠損値を多く含む属性は、その属性値の生起確率が小さくなるために発見されるルール中には出現しない。そこでそのような属性はあらかじめ排除した。「Retinopathy Score」や「Nephropathy Score」といった属性がこれに該当する。

○属性の統合

属性の中には、属性間で何らかの相関を持つ時系列属性が含まれる。例えば属性「網膜症有無」は、年度ごとに値がそれぞれ異なる属性として定義されている。これらすべての属性を使

用した場合、これらの属性間の陽な関係がルールとして導出される可能性が高いが、それらが有用な知識となる可能性は低い。そこで、これらの属性を、「1994年から2000年までの網膜症の有無」という1つの属性に統合した。

○離散属性値の集約

離散属性「網膜症有無」と「腎症有無」は多値の属性値をとる。これらは、相関ルール発見アルゴリズムを適用する際それぞれ別個の属性に分解される。これは属性数を増加させるため、ルール発見時の計算時間の増大につながる。そこで、このような多値属性値を二値に集約した。

前処理の結果として、実際にルール発見に使用する属性は22個(連続属性4個、離散属性18個)となった。

さらに、今回適用したアプリオリアルゴリズムは、2値の属性値を持つ離散属性しか扱うことができない。そこで、連続属性については、幾つかの「教師なし大域離散化」を用いて離散化を行う。「教師なし大域離散化」は、各事例が持つ目標概念に関する情報を用いずに、事例集合全体を使って属性間の依存性を考慮せずに各属性を離散化する手法である。今回は、属性値を等範囲で分割するEqual Width Intervalを用い、分割数は2とした。また離散属性は、各属性値についてYesとNoの二値をとる個別の属性に分割した。

c　導出ルール数の抑制

単純に相関ルールを適用すると、特に最小指示度を小さくした場合に非常に大量の相関ルールが導出され、しかもそのほとんどは、人間にとってあたりまえの規則を表現するルールであった。そこで、そのようなあたりまえのルールを機械的に排除し、導出ルール数を抑制するために、以下に述べる内部確信度を導入する。これは直感的には、相関ルールの条件部を構成する個々の属性は相関ルールの結論節と相関が見られないが、条件部全体が結論節と強い相関を持つルールのみを導出する操作に相当する。たとえば最小確信度を0.8でルール導出を試みて、

(1)　A ⇒ X　(確信度 0.82)

(2)　B ⇒ X　(確信度 0.81)

(3)　A,B ⇒ X　(確信度 0.80)

という結果が得られた場合、(3)は新しいルールとしての意味は乏しい。(1)、(2)の確信度より(3)の確信度が相当程度高い場合に意味がある。組み合わせで確信度が上昇することを表現する場合、割合で示すこともできるが、本研究では目標最小確信度より低い内部確信度を設定し、この例で言えば(1)、(2)は内部確信度より低く、(3)が目標最小確信度より高い場合だけに新し

表1 最小特定人数の計算例

患者.生年月日(年、月、日)	:30. 6人
患者.生年月日(年、月、日) ＋ 患者.性別	:15. 3人
患者.生年月日(年、月)	:368人
患者.生年月日(年、月) ＋ 患者.性別(女性)	:152人
患者.年齢(60歳)	:2万5千人
患者.住所(高槻市)	:12万8千人
患者.住所(高槻市安岡寺)	:4332人
患者.年齢(60歳)＋患者.性別(女性)	:1万1千人
患者.年齢(60歳)＋患者.性別(女性) ＋ 患者.住所(安岡寺)	:1170人
患者.年齢(60歳)＋患者.性別(女性) ＋ 保険適用.傷病.名称(胃がん)	:89人

いルールとして導出することとした。

すなわち、相関ルールR: Y1, Y2, … Ym ⇒ X に対して以下の相関ルールRk :Yk ⇒ X (1 ≦ k ≦ m)を考える。相関ルールRkは、ルールRの条件部に現れる属性を１つだけ条件部に含むルールである。このとき、confidence(R)をルールRの確信度とすると、

∀k[confidence(Rk) ＜ θ1 ∧ confidence(R) ＞ θ2]

であるようなすべてのルールを導出する。θ1を内部確信度と呼ぶ。θ2は最小確信度であり、θ1 ≦θ2である。

なお、本研究では倫理面の問題は存在しない。実診療情報を用いているが、完全に無名化されており、結果には個人を特定できる情報はまったく含まれていない。

C. 研究結果

１． データの無名化について

表１は生年月日の粒度別や特定の年齢、および特定の住所の最小特定人数および他の情報項目との組み合わせの最小特定人数を示す。

２．データマイニングおよび前処理について

前処理を施した糖尿病データベースに対し相関ルール発見手法を適用した。最小支持度を4、10、30、50％と変化させた。また最小確信度は90％とした。2.3節で述べた内部確信度としてそれぞれ、50、70、80、90％を用いた。実験結果

表2

最小確信度(%)	最小支持度(%)	内部確信度＜90%		内部確信度＜80%	
		ルール数	時間[sec]	ルール数	時間[sec]
90	50	115(87)	15.158	28(0)	14.949
	30	603(574)	109.843	29(0)	109.235
	20	1164(1134)	395.166	35(5)	395.141
	10	2193(2161)	2611.314	42(10)	2619.225
	4	3262(3229)	13297.554	54(21)	16203.318
最小確信度(%)	最小支持度(%)	内部確信度＜70%		内部確信度＜50%	
		ルール数	時間[sec]	ルール数	時間[sec]
90	50	28(0)	15.088	28(0)	14.912
	30	29(0)	110.584	29(0)	108.577
	20	30(0)	395.653	30(0)	394.558
	10	32(0)	2619.407	32(0)	3868.889
	4	34(1)	13272.743	33(0)	13334.230

表3

Coronary Disease (Yes), Hyperlipidemia (Yes), Lentinotomy (No) ⇒ Hypertension (Yes)
(支持度:4.96% / 確信度:91.18%)

Photocoagulation Therapy (Yes) ⇒ Retinopathy (not NDR)
(支持度:12.63% / 確信度:96.93%)

を表２に示す。前述した内部確信度を用いたルール数の抑制は、ルールのサイズ(条件部と帰結部に含まれる属性数)が2以上のルールに対してしか適用できない。表２中の括弧内の数字は、サイズが2以上のルール数、つまり実際に内部確信度による抑制が適用されたルールの数を示す。また得られたルールの例を表３に示す。

D. 考察

データの無名性に関して、表１は母集団として過去５年間に大阪医科大学付属病院で利用された患者基本情報３２万件とその関連データを用いたものであり、たとえば母集団が１万件

程度では生年月日を日まで特定すれば最小特定人数はほぼ１になることが予想される。また男女比はほぼ１対１と仮定した場合と実計算された最小特定人数はほぼ同じ値をしめしたが、年齢は対象地域（大阪府北摂地区で人口は約１１０万人）の年齢分布とは異なり、表には示していないが、６０歳代にピークを持つ。また住所も全国的に見れば大きな偏りを示すことは自明であり、計算例でも高槻市が半数近くを占めている。町名レベルでも大きな公的病院のある地域では最小特定人数が著しく低くなる傾向があった。

最小特定人数をデータの二次利用の匿名性の根拠として用いる場合、説明の対象は調査の対象となる人であり、たとえば疾患分布などの専門的知識を持たないと考えざるを得ない。したがって疾患分布で母集団に特異性があるかどうかは検討していない。このように母集団の偏りが存在し、説明の対象者に医学知識の存在を仮定できないために、今回おこなった検討では男女比を除いて、予測値を用いることはできないと考えられる。一方で最小特定人数の計算アルゴリズムは単純であり、正しく項目整理されたデータベースがあれば、簡単であり、実用に用いる点で大きな問題はない。

問題は項目整理であり、多施設間研究などでは項目の定義が異なれば、最小特定人数の計算はできない。もちろん多施設間研究そのもののために項目の同一定義は必要になり、その都度定義をそろえても理論的には計算可能であるが、データベースの再設計をする必要があり、また単純とはいえ、最小特定人数の計算アルゴリズムも毎回実装しなければならない。本研究ではJ-MIXを用いたが、これはきわめて有効であった。もちろんJ-MIXでは調査のための項目としては不足があることも考えられるが、かなり網羅的であり、少なくともJ-MIXを基本にすることで、調査のためのデータベースの設計や最小特定人数の計算アルゴリズムの実装は大幅に簡略化されると考えられる。

データマイニングについて、表２では最小支持率を低く設定すると、相関ルールでは著しい数のルールが導出される。本研究の最終的な目標は診療現場では気づかれにくい知識すなわちルールの検出であり、当然のことではあるが、そのルールが表現される事象の数は少ない。たとえば連続して受診した１００名の患者の中で、５０名に見られるようなルールは容易にその存在に気づく。したがって少ない例数の事象からルールを導出できなければ意味がない。したがって最小支持率は低く設定する必要がある。こ

の場合、問題になるのは既知のルールや、自明のルールの存在で、最終的には経験ある専門化が判断するか、教科書的な医学知識を備えた判定システムが必要になる。しかし、その前にデータマイニングの手法に内在した冗長性を排除することができれば、効率をあげることが可能になる。本研究では条件節に複数の項目が含まれる場合に内部確信度を導入し、条件節の項目が単一の場合よりも、確信度が一定以上高い場合だけをルールとして抽出することを試みた。この方法による導出ルール数の抑制は、抑制しない場合に比べて導出ルール数が約4分の１(内部確信度90％の時)に減少しており、導出ルール数だけ見るとうまく働いているように思われる。しかし、導出されたルールについての検証は行っておらず、診療根拠につながるような有用な知識が導出されているか、あるいは、逆に有効な知識となり得るルールを落としてしまっていないか、今後十分検証する必要がある。また表３の例の２つ目のように、条件節が単一項目からなる自明のルールはこの方法では排除できない。これについては知識データベースによる再解析が避けられないと考えられる。

E. 結論

無名性の指標として最小特定人数を用い、大阪医科大学付属病院のデータを用い、計算可能で有用なことを示した。診療情報を当該個人の健康回復や維持などの本来の目的以外に使用する場合、あらかじめ説明し同意を得ることが必要と考えられるが、まったく本人を特定できない場合で、公益目的であれば、その限りではない。また本人の特定がある程度困難な場合は同意を得やすいであろう。しかしながらどの程度困難であるか、定量化する方法はこれまで存在しなかった。本研究で用いた最小特定人数は他に考慮すべき要素はあるものの比較的簡単に計算することがで、良好な指標となることがわかった。

データマイニングに関しては、今年度は、相関ルール発見手法を適用し、導出ルール数を抑制することで、大量の診療データからなんらかの相関ルールを導出できることを確認した。しかし、導出されたルールが単なるパターンではなく、診療根拠につながりうる知識であるかどうかの検証はまだ行っていない。今後は、専門家による評価などにより、導出されたルールの検証を行う必要がある。相関ルールの導出に関しては、過剰に導出されるルールからより興味深いルールだけを導出するために、確信度や時系列の情報を用いたルール導出アルゴリズムの改良を検討しており、次年度はそれらの手法を

適用し有効性を検証する予定である。また、連続属性の離散化に関しては、今回用いた離散化手法は、データを離散化する閾値として診療データを反映した意味を持つ値を用いているわけではないため、よい方法とは言い難い。データの性質をより反映した離散化手法について検討する必要があると考えている。さらに、今回用いた糖尿病データベースに対し、教師付き学習である決定木学習を適用した結果、幾つかの目標概念については、簡潔な決定木を導出することができることを確認した。決定木学習は、利用者があらかじめ学習目標を事前に指定しなければならないため、本研究の目的である動的な診療根拠の生成に、それ単独で適用するのは難しい。しかし今回用いた相関ルール発見手法との組合せによって、決定木学習に対して学習目標をある程度機械的に設定し、学習を行うことが可能と考える。これは次年度の検討課題としたい。

F. 健康危険情報

なし。

G. 学会発表

山本隆一、増田剛、他：診療情報の無名性の定量化に関する研究、第２０回医療情報学連合、浜松、2000.

厚生科学研究費補助金(医療技術評価総合研究事業)

分担研究報告書

標準データ項目セットを用いた知的データベースによる診療根拠の動的生成に関する研究

— データ項目セットの整備と利用方法 —

分担研究者　大江 和彦　東京大学医学部附属病院中央医療情報部教授

研究要旨

大規模診療データベースから自動的に知識発見を行うためには、電子カルテシステムで使用される症状・所見コード（8項目）、診療問題コード(8項目)についての標準化作業が必要である。また、検体検査、放射線検査、生体検査、内視鏡検査、病理検査、細菌検査、超音波検査について、検査結果値の表記に関する標準化と検査結果実施記録項目セットの作成を行う必要がある。

A. 研究目的

筆者らは平成11年度に電子保存された診療録情報の交換のためのデータ項目セット The Japanese Set of Identifiers for Medical Record Information Exchange (J-MIX)の開発作業を担当した。これは、医療機関（診療所、病院など）において電子保存されている患者の診療データの一部または全部を、他の医療機関に電子的に送信する場面で、送信されることがあるデータ項目の一覧であり、各データ項目には、名称、データ型などの属性が付与されている。多くの既存の医療文書、医療情報交換用の規格、さまざまなカルテの記載用紙などで使用されているデータ項目が収集・整理、取捨選択された結果、診療録情報交換データ項目セットは最終的に１６１６項目からなる。

診療データベースから臨床的知見を抽出する場合、処理すべきロジックや得られた結果を流通可能な形式で表現できるようにしておくことが非常に重要であり、ここにJ-MIXが活用できると考えられる。そこで、診療データベースから臨床的知見を抽出する場合にJ-MIXを活用する上での問題点と今後改良すべき点を検討する。

B. 研究方法

臨床知見を発見する処理上必要となる項目の大部分はJ-MIXでコード型または構造型となっている。そこでコード型と構造型項目を抜き出し、それらを分類して今後必要となる臨床検査項目レベルについて、適切な粒度、他の既存のコードの利用方法について検討を行った。

C. 研究結果

コード型の項目は、３４６項目（21.4%）であり、それらが使用されている項目は下表のようであった。

	コード	コード体系コード	因子コード	因子コード体系コード	国コード	続柄コード	分類コード	分類コード体系コード	頻度コード	薬剤コード	薬剤コード体系	総計	分類
HB			2	2								4	検査
HC			2	2								4	検査
HIV			2	2								4	検査
MRSA			2	2								4	検査
感染症因子			2	2								4	検査
肯定検査名							1	1				2	検査
否定検査名							1	1				2	検査
結核菌			2	2								4	検査
検査	1	1					1	1				4	検査
梅毒			2	2								4	検査
家族歴疾患							2	2				4	疾患
鑑別診断	1	1					1	1				4	疾患
傷病	3	3										6	疾患
診断	4	4					4	4				16	疾患
総合診断							1	1				2	疾患
臨床診断名							2	2				4	疾患
家族						2						2	社会
自覚症状	1	1					1	1				4	症状
身体所見	1	1					1	1				4	所見
手術	2	2					2	2				8	処置
ID発行機関	23	23										46	組織
医療機関	34	34										68	組織
記録医療機関	1	1										2	組織
記録診療科	1	1										2	組織
主担当診療科	1	1										2	組織
住所					16							16	組織
傷病診療科	1	1										2	組織
診療科	28	28										56	組織
滞在病室名							2	2				4	組織
滞在病棟名							2	2				4	組織
退院後診療科	1	1										2	組織
担当医療機関	1	1										2	組織
担当診療科	1	1										2	組織
発行医療機関	3	3										6	組織
発行診療科	3	3										6	組織
計画							3	3				6	治療
治療							2	2				4	治療
傷病部位	3	3										6	部位
部位	4	4										8	部位
診療プロブレム	1	1					1	1				4	問題
現投与									1	1		2	薬剤
指示							1	1				2	薬剤
退院時投与									1	1		2	薬剤
入院時投与									1	1		2	薬剤
総計	119	119	14	14	16	2	28	28	3	3		346	

全体の21%をコード型がしめているが、整備すべきコードを分類してみると、検査コード（36項目）、疾患コード（36項目）、症状・所見コード（8項目）、手術処置コード（8項目）、組織コード（220項目）、治療コード(10項目)、部位コード（14項目）、診療問題コード(8項目)、薬剤コード（8項目）の9種類にまとめることができた。

次に構造型について分析する。構造型は28項目あり、指示に関するもの７項目、実施記録情報に関するもの１８項目、それ以外の文書情報３項目であった。今回の利用目的では、実施記録情報に関するもの１８項目、なかでも検査結果を取り扱う8項目（検体検査、放射線検査、生体検査、内視鏡検査、病理検査、細菌検査、超音

波検査、各種検査のそれぞれ実施記録情報）が重要であると考えられる。

D. 考察

データマインニングに必要となるデータ項目は、9種類のうち組織コードを除く8種類（126項目）であると考えられる。検査コードはすでに臨床病理学会項目コードが制定されておりいわゆる検査部検査であればそれが使用できる。疾患コードおよび部位コードは現在制定作業中の病名コード第2版を適用できる。また、手術処置コードおよび治療コード(10項目)は不十分ではあるもののMEDIS手術処置コードが適用できるであろう。さらに、薬剤コードもHOTコードが適用できる可能性が高い。そこで、残る症状・所見コード（8項目）、診療問題コード(8項目)について、コード作成とその実運用が必要である。

構造型に含まれている8種類の検査実施記録情報に記述される検査結果値のコード化およびデータ型の定義は本研究上極めて重要な問題である。これまで検査項目コードの標準化はいろいろな場面で論じられ、標準化が進んできたが、結果値に表記に関する標準化は未開拓の課題である。たとえば、ある検査項目の結果値が異常なしであるとき、「異常無し」「正常」「n.p」「W.N.L.」「－」などさまざな表記で記載されており、これらをいずれも同値としてコンピュータ上取り扱うには標準化が不可欠である。

E. 結論

大規模診療データベースから自動的に知識発見を行うためには、電子カルテシステムで使用される症状・所見コード（8項目）、診療問題コード(8項目)についての標準化作業が必要である。また、検体検査、放射線検査、生体検査、内視鏡検査、病理検査、細菌検査、超音波検査について、検査結果値の表記に関する標準化と検査結果実施記録項目セットの作成を行う必要がある。

F. 研究発表

学会発表

波多野賢二、大江和彦、山本隆一他：電子保存された診療録情報の交換のためのデータ項目セットの開発、第 20 回医療情報学連合大会、浜松、2000.

平成 12 年度厚生科学研究費補助金（医療技術評価総合研究事業）
分担研究報告書

標準データ項目セットを用いた知的データベースによる診療根拠の動的生成に関する研究

患者プライバシを保護したデータベースアーキテクチャの研究

分担研究者　坂本　憲広　九州大学医学部附属病院　講師

研究要旨

保健医療分野において、高品質の医療の実現を目指して、EBM の研究実践がなされている。一方、社会の様々な分野における情報化に伴い、保健福祉医療分野においても多くの情報システムが導入され、種々の情報サービスが提供されるようになってきている。こうした保健福祉医療情報システムの普及に伴い、患者に関する多種多様な保健医療情報が収集され、集積され、大規模な保健医療情報データベースとして管理されるようになってきている。こうした大量の情報を有効に活用し、EBM のための“根拠”を効率的に導出するためには、データマイニングあるいはデータベースからの知識発見と呼ばれる、人工知能技術が有効であると考えられる。

データベースからの知識発見では、統計解析とは異なり、発見目標である知識が何であるかは前提とされていない。そのため、発見に至る過程においては、データベースに格納されているあらゆる項目のあるゆるデータについて、機械的（自動的）に集計、分析を繰り返す。従って、解析対象であるデータベースが、患者の氏名や住所などの個人を識別可能な情報を含んでいる場合、これらの情報も解析対象として処理され、時には、個人が識別可能な状態で研究者に解析結果として示される可能性がある。こうした状況は、患者プライバシの保護の観点から非常に重大な問題であると考えられる。

そこで、本研究では、患者プライバシを保護しながら、データベースの知識発見を行うことを可能にする、データベースアーキテクチャの構築を目指す。

A. 研究目的

保健医療分野において、高品質の医療の実現を目指して、EBM の研究実践がなされている。一方、社会の様々な分野における情報化に伴い、保健福祉医療分野においても多くの情報システムが導入され、種々の情報サービスが提供されるようになってきている。こうした保健福祉医療情報システムの普及に伴い、患者に関する多種多様な保健医療情報が収集され、集積され、大規模な保健医療情報データベースとして管理されるようになってきている。こうした大量の情報を有効に活用し、EBM のための“根拠”を効率的に導出するためには、データマイニングあるいはデータベースからの知識発見と呼ばれる、人工知能技術が有効であると考えられる。

しかしながら、患者の実データを用いて臨床研究や解析を行うに際しては、常に倫理面に配慮し、患者プライバシに留意しなければならない。

本研究では、患者プライバシを保護したまま、臨床的に有益な知識をデータベースから発見する手法について研究を行う。

B. 研究方法

データベースからの知識発見では、統計解析とは異なり、発見目標である知識が何であるかは前提とされていない。そのため、発見に至る過程においては、データベースに格納されているあらゆる項目のあるゆるデータについて、機械的（自動的）に集計、分析を繰り返す。従って、解析対象であるデータベースが、患者の氏名や住所などの個人を識別可能な情報を含んでいる場合、これらの情報も解析対象として処理され、時には、個人が識別可能な状態で研究者に解析結果として示される可能性がある。こうした状況は、患者プライバシの保護の観点から非常に重大な問題であると考えられる。

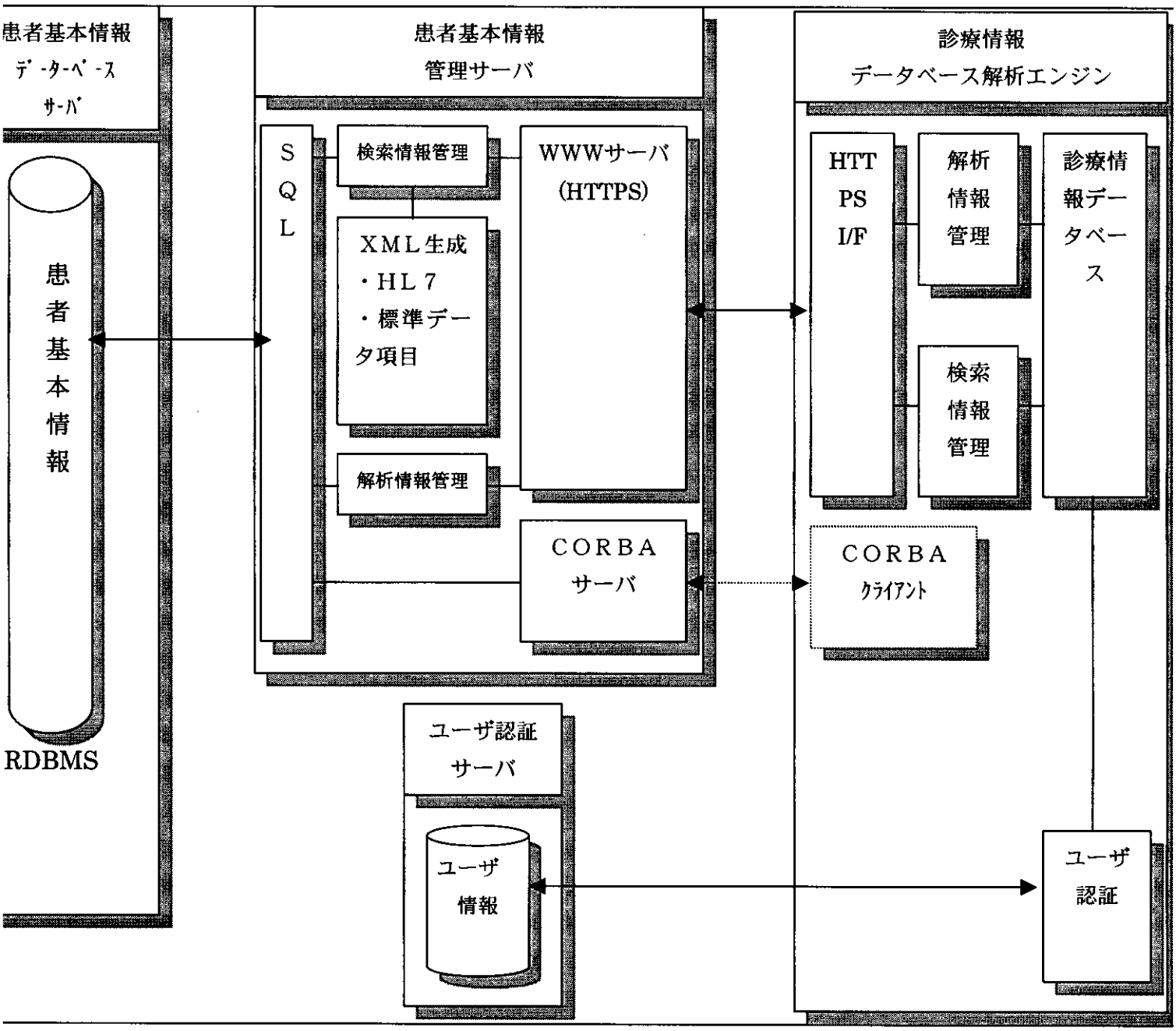
そこで、本研究では、患者プライバシを保護しながら、データベースの知識発見を行うことを可能にするため、患者プライバシに強く関連する患者基本情報を管理するデータベース（患者基本情報管理データベース）と臨床医学的データを管理する診療データベースを分離し、必要な際のみに、必要最小限の匿名化データを患者基本情報管理データベースから解析エンジンに伝達する、データベースアーキテクチャを提案する。

（倫理面への配慮）　本研究は、データベースの構造あるいは構成そのものを対象とした研究であり、実際の患者あるいは患者情報を対象とした研究ではないため、本研究の遂行において、特に倫理的な問題が関連することはないと予想される。

C. 研究結果

システム設計

患者基本情報管理データベースを診療情報データベースから分離し、その間の通信インターフェイスなどの



システム設計を図の通り行った。

システム間の通信はHTTPSを用いて暗号化される。また、どのようなデータを検索したかの情報（検索情報）とその検索結果をどのような解析に用いたかの情報（解析情報）は、それぞれのデータベースシステムにおいて管理しており、より厳密なプライバシ保護の実現を目的としたセキュリティ監査を可能とするシステム設計となっている。

患者基本情報の要求は、標準データ項目セットに準拠した内容で可能であり、患者基本情報管理サーバーがこれをSQLに変換し、患者基本情報データベースより情報を抽出する。また、

結果は匿名化され、標準データ項目セットの XML 文書として返信される。匿名化の手法に関しては、今後の研究課題である。

データベース構造設計

データベースの構造は今後の拡張性を考慮して、HL7RIM に準拠した。そこで、本研究では、標準データ項目セット（J-MIX）のうち、患者基本情報に関連した部分を HL7RIM にマッピングした。その結果の一部を書きに示す。

J－MIX	HL7ver3.0 RIM
患者.ID	Role-Entity<id>
患者.ID 発行機関.名称	Role-Entity<id>
患者.ID 発行機関.ｺｰﾄﾞ	Role-Entity<id>
患者.ID 発行機関.ｺｰﾄﾞ体系ｺｰﾄﾞ	Role-Entity-Role-Role_relationship-Role-Entity<id>
患者.氏名	Role-Entity-Entity_name<nm>
患者.姓	Role-Entity-Entity_name<nm>
患者.名	Role-Entity-Entity_name<nm>
患者.カナ氏名	Role-Entity-Entity_name<nm>
患者.カナ姓	Role-Entity-Entity_name<nm>
患者.カナ名	Role-Entity-Entity_name<nm>
患者.生年月日	Role-Entity-Living_subject<birth_dttm>
患者.性別	Role-Entity-Living_subject<administrative_gender_cd>
患者.年齢	Role-Entity-Living_subject<birth_dttm>
患者.職業	Role-Entity-Role-Role_relationship-Employee_Employer<job_cd>
患者.住所	Role-Entity-Living_subject-Person<addr>
患者.住所.国ｺｰﾄﾞ	Role-Entity-Living_subject-Person<addr>
患者.住所.都道府県	Role-Entity-Living_subject-Person<addr>
患者.住所.市区部名	Role-Entity-Living_subject-Person<addr>
患者.郵便番号	Role-Entity-Living_subject-Person<addr>
患者.電話番号	Role-Entity<telecom>
患者.緊急連絡先	Role-Entity<telecom>
患者.FAX 番号	Role-Entity<telecom>
患者.電子ﾒｰﾙｱﾄﾞﾚｽ	Role-Entity<telecom>
患者勤務先.名称	Role-Entity-Role-Role_relationship-Role-Entity-Organization<org_nm>
患者勤務先.住所	Role-Entity-Role-Role_relationship-Employee_Employer<addr>
患者勤務先.住所.国ｺｰﾄﾞ	Role-Entity-Role-Role_relationship-Employee_Employer<addr>
患者勤務先.住所.都道府県	Role-Entity-Role-Role_relationship-Employee_Employer<addr>
患者勤務先.住所.市区部名	Role-Entity-Role-Role_relationship-Employee_Employer<addr>

今年度は、本研究の初年度であり、本研究が最終目標とする、患者基本情報

と診療情報の分離について、システム面および情報構造の観点から、そのフィージビリティを検証した。
データベースを分離すことにより、患者プライバシの保護が実現されることは明らかであるが、データベースからの知識発見のための解析処理に際して、データベース間の情報通信が発生し、解析速度が低下することが懸念される。しかしながら、患者基本情報は、有意義な医学的知識を発見する上において、一般的にそれほど重要ではないと考えられる。そのため、患者基本情報を除く、診療情報のみを解析対象とすることで、探索空間が減少し、解析速度の向上と解析結果の精度の向上が期待される。この点については、計算量からの理論的アプローチと同時に、プロトタイプシステムを利用した実証実験が必要であろう。

E. 結論

本年度の研究により、患者基本情報を診療情報から分離した、データベースアーキテクチャが実現可能であることが示された。来年度以降は、提案システムのプロトタイプを研究開発し、このデータベースアーキテクチャが、患者プライバシの保護を実用的に実現できることを示す。また、同時に、診療根拠を動的に生成するためのデータベースからの知識発見の処理に際して、このデータベースアーキテクチャが、十分な処理性能を実現可能であることを、理論的および実験的に検証していく予定である。

F. 研究発表

これまでのところは本研究に関する研究発表は行っていない。

G. 知的所有権の取得状況

なし