化で個人情報とみなさなくてよいかは社会的なコンセンサスが必要で、当面は説明と同意の原則に従う必要がある。今回、プライバシー保護ガイドラインを試作したが、医療のすべての場面を網羅しているとは言えず、さらに完成度を上げる必要がある。そして今後完成したガイドラインにしたがって診療情報の収集を行い、研究を進める必要がある。

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。

いずれにしてもこの結果は静的なデータベースに対して知識発見を試みたもので、相関ルール発見手法が本研究で目指す、動的な診療根拠の発見に有用であることは示せたものの、急性で流行性または集団発生の疾患において動的なデータマイニング行う場合やまれな状況についての知識発見については、単純な相関ルール発見手法の適応やデータの前処理だけでは適応できない。増田が提唱したフレームワークはユースケースを限定した知識発見のフレームワーク

で今後この方向で研究を進めていくことによって動的な診療根拠生成に到達することが可能になったと考えられる。

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。

PC-UNIXサーバを用いて廉価かつ用意に実験用VPN環境を構築することができた。MS-Windows環境ではさらに工夫が必要であるが、実現可能と考えられる。

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。

さらに診療根拠の動的生成のユースケースを１．急性で流行性のまたは集団発生をする疾患の場合、２．低頻度の薬剤副作用のようにまれで個々の診療現場では気づきにくい状況、という２つに限定し、その解決のためのフレームワークを示した。

F．健康危険情報

なし。

G．論文・学会発表

著書・論文

１．医療経済研究機構監修、医療白書 2001 年度版（プライバシー保護としての医療情報のセキュリティ対策　山本隆一）、日本医療企画、東京、2001

２．財団法人四国産業・技術振興センター編、電子カルテネットワーク（診療情報交換とセキュリティ、電子カルテネットワークの技術的課題－セキュリティ山本隆一）、エムイー振興協会、東京、2001

３．山本隆一、医療情報のセキュリティ、システム／制御／情報、44、576-582、2000

４．山本隆一、電子保存新基準について　－運用規定策定の試みと評価－、映像情報、32(2)、92-96、2001

５．山本隆一、医療情報システムのセキュリティモデル、医学のあゆみ、196、277-281、2001

６．山本隆一、ネットワーク時代の身分証明と安全性確保 － 電子化された診療情報のセキュリティについて －、治療、83、245-251、2001

７．山本隆一、ネットワーク時代の医療情報の安全性、BIO Clinica、16、721-725、2001

８．山本隆一、増田　剛、濱田松治、生体識別（Biometrics）、Innervision、16(7)、14-16、2001

９．山本隆一、医療情報のセキュリティ、Mebio、18(5)、132-138、2001

１０．K.Hatano, K.Ohe, R.Yamamoto: Development of the set of data identifiers for medical record information exchange. Proceedings of MEDINFO2001, V.Patel etal.(Eds), Amsterdam:IOS press, 706, 2001.

１１．M. Sproger, P. Kokol, M. Zorman, V. Podgorelec, R. Yamamoto, G. Masuda, N. Sakamoto: Supporting Medical Dicision with Vector Dicision Trees. Proceedings of MEDINFO2001, V.Patel etal.(Eds), Amsterdam:IOS press, 552, 2001

発表

１．山本隆一、シンポジウムＨＩＰＡＡの動向　HIPAA 関連規則、特に Security および Privacy 保護規則に関する研究、第 21 回医療情報学連合大会、東京、2001

２．R. Yamamoto, Practical Strategies for Addressing the General Privacy Act in Japan and OECD Privacy Guidelines. 3rd China-Japan-Korea Joint Symposium on Medical Informatics, Tokyo, Japan, 2001

３．増田 剛，山本隆一：相関ルール発見手法を用いた診療データベースからの知識発見における導出ルール数の抑制，第 21 回医療情報学連合大会，pp. 466-467，2001 年 11 月．

H. 知的財産権の取得状況

なし

厚生科学研究費補助金（21 世紀型医療開拓推進研究事業）
分担研究報告書

データマイニングを用いた動的な診療根拠の生成システムのフレームワーク構築に関する研究

分担研究者　増田　剛　大阪医科大学　病院医療情報部　助手

研究要旨

本研究では、通常の EBM の適用が困難なために動的な診療根拠の生成が必要となる二つのユースケースを考え、データマイニングを用いた動的な診療根拠の生成システムのフレームワークを構築した。一つ目のユースケースである、流行性・一過性の疾患に対する診療根拠の動的な生成に関しては、データマイニング手法の一つである相関ルールを用いたルール発見、導出されたルールのルールベースによる管理、意思決定を支援するための問合せ機構によるルールベースの後解析、ユーザへの効果的なルール提示のためのルール削減・要約手法を用いた枠組みを提案する。二つ目のユースケースである、薬品の低頻度の副作用に対する診療根拠の導出に関しては、相関ルール発見の枠組みにおいて薬品の低頻度の副作用を表すルールの性質を考え、ルールの確信度の差に基づくパターンの発見を用いる。本研究により、動的な診療根拠の生成システムに必要な要件を定義することができ、この枠組みに対して適用可能なデータマイニング手法とその問題点を明らかにすることができた。

A. 研究目的

近年、より質の高い保険医療サービスを提供するために Evidence Based Medicine (EBM)の実践が重要な考え方となっている。しかし、一般に EBM は過去の文献で示された診療根拠に基づくため、そのような文献による診療根拠を待つことができない疾患への適用は難しい。また、例数が極めて少なくこれまでに診療根拠が確立されていない状況での適用も困難である。一方で、保険医療情報の電子化が進み様々な医療情報システムが保険医療分野に導入されてきている。また、ネットワーク指向の電子カルテシステムの登場と診療情報の標準化技術の発展により、診療情報を広範囲から収集し統一的に扱うことが可能になりつつある。これに伴い患者に関する様々な保険医療情報が収集・蓄積され、大規模な保険医療情報データベースが形成されるようになってきている。これらの大量の情報を機械

的に解析することにより、動的に診療根拠を導出することができれば、EBM の適用が難しい状況において、EBM を実践した質の高い保険医療サービスの提供が可能となる。データマイニング技術は、このような大量に蓄積されたデータから診療根拠となり得るパターンを動的に導出するための有効な手段の一つである。

データマイニングに関しては、これまで情報工学や知識工学の分野で研究されてきた。その中には対象として医療データを扱った研究も数多くなされている。しかし、広域的に収集された診療情報から、診療根拠を動的に生成する研究に関して、どの種のデータマイニング技術をどのように活用していくべきなのか、その活用方法は未だ明らかになってはいない。そこで本研究では、データマイニングを用いて診療根拠を動的に生成するシステムのためのフレームワークを構築することを目的とする。

前年度までの検討で明らかなように、本システムのフレームワークの構築には、診療情報のプライバシ保護のためのセキュリティ技術や広域的なデータ収集技術に関する議論が不可欠である。しかし、本研究ではデータマイニング手法のみに着目してフレームワークの構築を行なう。

また本研究では、診療データを動的に解析することによって得た診療根拠を用いた EBM を、Dynamic EBM (DEBM)と定義し、本研究で提案する動的な診療根拠生成システムを DEBM システムと呼ぶ。

B. 研究方法

一般に、ある問題領域に対してデータマイニング手法を適用する場合、データマイニングの目標を定めることが必要となる。解析目標を定めずにデータマイニング手法を無秩序に適用しても有効な知識を得ることは難しい。そこで、本研究では診療根拠の動的な生成が必要となる状況として、次の二つのユースケース（システムの利用例）を考える。それらのユースケースに適した DEBM システムのフレームワークの構築を目指す。フレームワークの構築に際しては、情報工学や人工知能の分野でこれまで研究されてきたデータマイニング手法を調査検討し、適用可能な手法やその適用方法、既存の手法の問題点を明らかにする。

ユースケース１：流行性・一過性の疾患

流行性・一過性の疾患は、過去の文献に示された診療根拠を得ることができないため、従来の EBM を適用することが困難である。このような状況では、文献に頼ることができないため、広域的に収集され日々蓄積されている診療データを動的に解析することによって、診療根拠の手がかりとなる関係を導出することが必要となる。これと同時に、診療時に医師がそれらの解析結果を分析し意思決定を行なうための支援機能が必要である。例えば、今年のインフルエンザにある薬品が有効であるか、あるいは、非常にまれな集団食中毒に対してある抗生物質が有効かどうかを検証するといった場合が考えられる。

ユースケース２：薬品の低頻度の副作用

薬の低頻度の副作用は、臨床現場の医師にとって非常に気付きにくい関係であり、そ

れを人手で発見するためには相当な例数の解析が必要となる。また、過去に確立された診療根拠がないため、従来の EBM を適用することは難しい。そのため広域的に収集されたデータを動的に解析することによって薬品の副作用と疑わしい関係をできるだけ迅速に機械的に見つけ出す必要がある。

以上２つのユースケースを本研究での DEBM システムに対する要件と考え、システムのフレームワークを検討する。

C. 研究結果

C.1. ユースケース１に関する検討

このユースケースの場合、データマイニングの解析目標は、必ずしも医学的に新しい知識や興味深い知識の発見というわけではない。確率的に期待される値以上に何らかの関連があるパターンを網羅的に発見することが重要となる。またデータマイニング手法によって発見されたパターンに対して医師が分析を行ない診療時の意思決定を行なうことができるような導出ルールの分析機能が重要となる。そこで本研究では、データマイニング手法を適用した結果に対する後解析(Post-analysis)アプローチを採用する。これは、データマイニング手法を用いて事前にデータを解析し、導出されたパターンに対する後処理でユーザが求める診療根拠を導出するという二段階の解析を用いる方法である。

本研究では、データマイニング手法としてルール発見手法の一つである相関ルール発見手法を用いる。相関ルール発見手法は、後に述べる最小支持度と最小確信度というユーザが与える制約の元で網羅的にルールの探索を行なう。また、導出される相関ルールは直感的であり、さらなる解析へのフィードバックとして容易に用いることができる。さらに、並列・分散化への拡張が可能で大量のデータを扱うことができる。

相関ルール発見は次のように形式化できる。図１のようなトランザクションデータベース D を考える。$I=\{i_1,i_2,...i_m\}$ をアイテムの集合とする。このとき k 個のアイテムの組合せを長さ k のアイテム集合と呼ぶ。D 中の各トランザクション T は、$T \subseteq I$ であるようなアイテムの集合である。このとき相関ルールは $X \Rightarrow Y$ で表され，$X \subset I$，$Y \subset I$，$X \cap Y = \phi$ である。相関ルールはルールが持つ確信度（confidence）と支持度（support）の２つの指標によりそのルールの有効性を評価することができる。相関ルール $X \Rightarrow Y$ の確信度は、D 中で X を含むトランザクションのうち X と Y を共に含むトランザクションの割合として定義され、そのルールの正確性を評価する。一方、相関ルールの支持度は、D 中で $X \cup Y$ であるトランザクションの割合として定義される。支持度は、その相関ルールが適用可能なデータ数を表しており、値が大きければ大きいほど一般性が高いルールとみなすことができる。相関ルール発見は、あるトランザクション集合 D が与えられたとき、ユーザが指定する最小支持度と最小確信度を満足するすべての相関ルールを発見する問題と定義できる。これは、各属性が「網膜症有無」のような離散的な値をとる離散属性と、「HbA1c」のような連続値をとる連続属性から構成される、図２上に示すようなデータベースに対して容易に発展可

能である。連続属性は前処理により離散化される必要がある。各離散属性について「属性＝値」の組を一つのアイテムとし、値としてYesをとる属性値の集合をトランザクションとすることにより、図1に示すアイテム集合とトランザクションデータベースを作ることができる。このデータベースから得られる相関ルールの例を図2下に示す。

後解析の手法としては、本研究では問い合わせ指向のアプローチを採用する。これは導出されたルールをルールベースとして管理し、そのルールベースに対する質問処理によって、ユーザが対話的にルールを解析することを可能にする。データマイニングに関するこれまでの研究で、ルールの興味深さについて、客観的な評価指標や、例えば意外性といった主観的な評価指標についての研究が行われてきた。しかし、EBMのための診療根拠の導出においては、状況によって様々な要求が起こり得るため選択基準として唯一の基準を事前に設けることは難しい。それゆえ、ルール発見手法にこのようなルールの評価基準を導入するのではなく、本研究で採用した、後処理の中で対話的にルールの評価を行なうアプローチが有効であると考える。

後処理による対話的な導出ルールの選択や削減、枝刈りに関する研究としては、ユーザが明示するテンプレートに基づきルールを選択する手法や、異なる詳細度で表現される知識に対し、その知識に該当するルール、あるいはその知識に反するルールをそれぞれ提示する手法が提案されている。また、個々のルールに対する評価値の時間的な推移に関する問い合わせを用いてルールを選択する手法が研究されている。

本研究では初期的な実験として、ルールベースとルールの後解析ツールのプロトタイプを実装し、実際の診療データを使用して評価を行なった。ルールベースは、関係データベースとして実装し、相関ルール発見手法によって導出された相関ルールをルールテーブル、アイテムセットテーブル、アイテムテーブルの三つのテーブルを使って格納する。ルールベースへの問合せ手法には、SQLに基づく問合せ機構を採用した。前年度の研究で使用した糖尿病データベース 1251 事例に対して相関ルール発見手法を適用した。得られた約25万個の相関ルールをルールベースに格納し、後解析ツールにより解析を行なった。今回実験を行なった解析例とその応答時間を図3に示す。なおルールベースの関係データベース管理システムには PostgreSQL を使用し、CPU: Pentium3 600MHz、メモリ 384MB の計算機上で動作させた。また後解析ツールは Java アプリケーションとして実現し、データベースサーバと同一計算機上で実行した。

C.2. ユースケース2に関する検討

薬品の低頻度の副作用は、診療現場で人が気付きにくい、出現頻度が低いけれども意味のあるパターンである。ユースケース1で述べた相関ルール発見手法においてこのようなパターンは以下の性質を持つと考えることができる。

次の二つの相関ルール r と r' を考える。

$r: X_1, X_2, \ldots, X_k \Rightarrow Y$

$r': X_1, X_2, \ldots, X_k, X_{k+1} \Rightarrow Y$

ルール r' は、ルール r と同じ結論を持ち、条件部に付加的な属性を一つ持つ。このとき、(1)両者の指示度、確信度は共に極めて

低い、かつ、(2)両者の相関ルールの確信度の差が極めて大きい、つまり、確信度をそれぞれ$Conf(r)$、$Conf(r')$とすると、

$$|Conf(r') - Conf(r)| >> 0$$

が成り立つ。上記(1)、(2)の条件を満たす相関ルールの組において、r'の条件部の付加的な属性X_{k+1}は、結論部の属性Yにとってルールの確信度を変化させる重要な要素となる可能性があると考えられる。薬品の副作用という文脈においては、この付加的な属性が薬品の投与を表す属性に対応する。このようなルールの組を見つけることにより低頻度の副作用を表すパターンを発見することが可能である。

しかし、これまで述べた相関ルール発見手法を単純に適用しただけでは、このようなパターンは最小支持度の制約により発見できない。逆に最小支持度を極端に小さな値に設定すると、出現頻度の高い属性の組合せ爆発により無意味なルールが大量に導出されてしまう。また、相関ルールの支持度とは異なり、確信度の性質から探索空間の枝刈りを行なうことができない。このようなルールの組を効率良く発見するためには何らかの工夫が必要となる。この問題に対して、例外的なルールを常識ルールと例外ルールの組として導出する例外ルールの発見の応用が考えられる。薬品の投与を表す常識ルールとそれに対する例外ルールを見つけることで低頻度の副作用を表す可能性があるルールを発見することができる。

C.3. 導出されたルールの削減と要約

相関ルール発見手法は、ユーザが指定した支持度と確信度の制約を満たすすべてのルールを導出する。しかしその網羅性ゆえに結果として大量のルールが導出されるという問題がある。そして大量の自明で冗長なルールを含む場合が多い。そのため大量に導出されるルールからユーザにとって興味のあるルールを見つけることが非常に重要となる。本研究ではフレームワーク１において、ルールを網羅的に導出するため極めて大量のルールが生成される。それらを後解析において効率良く処理するために、ルールの削減・枝刈りや要約手法が必要となる。

後処理におけるルールの削減・要約手法として既存の研究で提案されている手法としては、相関ルール発見により導出されるルールを、一般ルール、要約、例外の三つ組みで整理することにより効果的にルールを提示する手法や、複数のルールの組合せから容易に推測できるルールを排除することでルール集合を要約する Direction Setting ルール手法が提案されており、導出された大量のルールの整理・評価に適用可能である。

我々は、昨年度から継続して行なってきた研究でルールフィルタリングによるルールの削減手法と枝刈り手法について検討を行ない、実際の医療データに対して適用し評価を行なった。以下に適用した手法についてそれぞれ述べる。

Filter 1:

$$R: Y_1, Y_2, ..., Y_m \Rightarrow X$$
$$R_k: Y_k \Rightarrow X (1 \le k \le m)$$

Rをm個$(m \ge 2)$の属性をルールの条件

部に持つ相関ルールとする。また、R_kをルールの結論部にRと同じ属性を持ち条件部にRのk番目$(1 \le k \le m)$の属性を持つ相関ルールとする。このとき、次の条件を満たすルールのみを採用する。

$Conf(R_k) < \theta \wedge Conf(R) \ge Minconf$

θはユーザから与えられた閾値であり、$0 \le \theta \le 1$である。またMinconfは最小確信度である。これはルールの条件部の個々の属性は結論部の属性と関連が低いが、それらの組合せを考えると結論部の属性と強い関連を持つルールのみを選択することを意味する。

Filter 2:

$R_m : Y_1, Y_2, ..., Y_m \Rightarrow X$

$R_{m+1} : Y_1, Y_2, ..., Y_m, Y_{m+1} \Rightarrow X$

R_mを条件部にm個$(m \ge 1)$の属性を持つ相関ルールとする。R_{m+1}を、ルールの結論部にR_mと同じ属性を持ち、かつルールの条件部に、R_mの条件部の属性に加え一つの属性を持つ相関ルールとする。このとき、R_{m+1}の確信度がR_mの確信度よりも大きい場合にのみ、相関ルールR_{m+1}を採用する。

ルールの枝刈り:

$R_k : Y_1, Y_2, ..., Y_k \Rightarrow X$

$R_m : Y_1, Y_2, ..., Y_k, ..., Y_m \Rightarrow X$

R_mをm個の属性を条件部に持つ相関ルールとし、R_k $(1 \le k < m)$を、条件部にR_mの条件部の属性の部分集合を持つ相関ルールとする。このとき、もしR_kの確信度がR_mの確信度よりも大きい場合、ルールR_mをR_kで置き換える。これは、ある事象を説明するルールが複数存在するときに、それらが同程度に正しければより簡潔なルールの方が望ましいという最小記述長原理に基づく枝刈りである。もしこの枝刈り処理によってルール集合の中に同じルールが二つ以上生成される場合は、それらのルールは一つのルールにまとめられる。

実験には、昨年度から使用している糖尿病データベースと、KDDチャレンジ2000と呼ばれるデータマイニング手法のコンテストで使用された髄膜炎データベースを用いた。それぞれのデータベースの説明を表1に示す。最小確信度を90%に固定し最小支持度を変化させ、単純な相関ルール発見手法の適用により導出されるルール数と、削減手法を適用した場合の導出ルール数を比較した。なおFilter1での閾値は90%とした。実験結果を表2に示す。結果より、二つの削減手法については、ルールの削減率だけみるとFilter1の方がFilter2と比べてより大きな削減率を示した。しかし、Filter1ではルールが過度に削減される傾向があるため、ユースケース1において後解析処理の中で用いる場合にはFilter2の方が適していると考える。また枝刈り手法については、手法の適用によってルール数が削減されると同時に平均ルール長が減少した。今回の実験では、ルール長さが4以下のルールのみを導出したが、より長いルールが生成される状況では、枝刈りの効果がより大きいと予想される。

C.4. 時系列データの解析

日々蓄積される診療情報から動的に意味のある診療根拠を導出するためには、時系列データの解析が必須である。これまで時系列データの解析に関して、ルール発見手法に時間的な概念を導入した手法が提案さ

れている。また別の方法として、時系列データの処理を前処理の問題に帰着させるアプローチがある。

我々は後者のアプローチを採用する。時系列属性を前処理の段階で集約することにより新たな属性を追加する。それらの変換されたデータベースに対し通常のルール発見手法を適用する。診療情報は日々収集され蓄積され続けるため、この処理を定期的に行なう必要がある。一般にある属性が持つ時間的な関係は数多く存在する。そのためそれらのすべての関係を前処理によって導出することは現実的ではない。しかし、特に EBM においては診療根拠として疾患の発症点からの時間的な変化が重要であることから、時間的関係を限定することにより前処理によるアプローチも実現可能である。このアプローチは、適用するルール発見手法に依存しないため、問題に応じて種々の手法を適用可能である。

また、相関ルール発見手法においては、動的に蓄積されるデータベースへに対して全体を再解析することなしに効率的にルールを発見する手法が提案されており、これらの手法の適用も可能である。

C.5. 全体のフレームワーク

以上の議論に基づき設計した DEBM システムのフレームワークを図 4 に示す。電子カルテシステムによって広域的に収集された診療データは、データベースに蓄積される前に前処理として時系列属性の処理が行なわれる。ルールマイニングエンジンは、ルール発見手法を適用することにより、前処理後の診療データからルールを導出する。このときルール発見手法は、例えば一日一回というように定期的に適用される。導出されたルールは、ルールベースに格納されてさらなる後解析のために管理される。ユーザは、ルール解析ツールを使用してルールベースに対して問合せを発行することで、流行性・一過性の疾患に対する診療根拠の導出を行なう。また、薬品の低頻度の副作用の発見については、前処理後の診療データベースに対して定期的にパターン発見手法が適用される。システムは、そのようなパターンが見つかった段階でユーザからの問合せを待たずに警告を出しユーザに知らせる。

D. 考察

D.1. ルールの後解析について

本研究で実装した後解析ツールのプロトタイプでは、属性の有無による単純なルールの選択機能しか実現していない。そのため導出された大量のルールを解析するには十分ではなく今後の改良が必要である。また、単純な問合せによっては、大量のルールが導出されてしまうため、本文で述べたようなルールの枝刈り手法や要約手法の適用が不可欠である。さらに、今回は相関ルール発見手法一回の適用で導出したルールベースを用いたため、ルール数は約 25 万件であった。しかし実際には、ルール発見手法を繰り返し適用するため非常に大量のルールが生成される。そのため、そのような大量のルールの全てを今回の実験のように一つのルールテーブルで管理するのではなく、例えばルールに含まれる属性の種類に応じて、ルール集合をある程度事前に決めたカテゴリに分類しておくことが、より効

率的な質問応答処理を実現するために必要になる。

さらにユーザが一旦行なった問合せを管理しておくことによって、ユーザが関心を持つ問合せを定期的に調べ、その問合せに関してルールベースの中で重要な変化が生じた場合にはそれを自動的にユーザに通知することは有用である。これは次年度以降に検討したい。

D.2. ルールの削減・枝刈りについて

本研究で行なった実験結果から、適用したフィルタリング手法によって冗長なルールが削減されるかどうかを考察した。実験に使用した糖尿病データベースから、

末梢神経障害 無 ⇒ 自律神経障害 無 (支持度 65.8% / 確信度 97.5%)

といった極めて自明なルールが得られた。その結果、データベース中で出現頻度の高い属性がこのルールの条件部に加わった冗長といえるルールが大量に導出された。この例の場合、そのようなルールは 785 個導出されたが、削減手法を適用した結果、Filter1 を適用した結果 1 個に、また Filter2 を適用した結果 332 個にまで減少した。同様に、

白内障 無 ⇒ 水晶体摘出 無 (支持度 73.1% / 確信度 96.6%)

という自明なルールに対して出現頻度の高い属性が加わったルールが 846 個導出された。しかし削減手法の適用により Filter1 で 1 個、Filter2 で 229 個に減少した。これらの結果から、本研究で提案した削減手法は冗長なルールの除去に対して効果があると言える。

E. 結論

本年度は、動的な診療根拠の生成が必要となる二つのユースケースを考え、データマイニングを用いた DEBM システムのためのフレームワークを構築した。一つ目のユースケースである、流行性・一過性の疾患に対する診療根拠の生成に関しては、データマイニング手法の一つである相関ルールを用いたルール発見と発見されたルールの後処理による解析により実現する。二つ目のユースケースである、薬品の低頻度の副作用に対する診療根拠の導出に関しては、ルールの確信度の差に着目したパターンの発見が有効であると考えた。

これらの検討をもとに次年度は、DEBM システムのプロトタイプを作成し、構築したフレームワークの有効性を評価する予定である。例えばこれまで実際に発生した薬品の副作用データや流行性の感染症データベースに対してシステムを適用することが必要であると考えている。また、時系列属性の前処理や薬品の副作用を表すパターン発見といった、さらなる検討が必要な課題について次年度も引き続き研究を行なう予定である。

F. 健康危険情報

なし

G. 研究発表

1．論文発表

なし

2．学会発表

増田 剛, 山本隆一: 相関ルール発見手法を用いた診療データベースからの知識発見における導出ルール数の抑制, 第 21 回医療

情報学連合大会， pp. 466-467， 2001 年 11 月.

H. 知的財産権の取得状況

なし

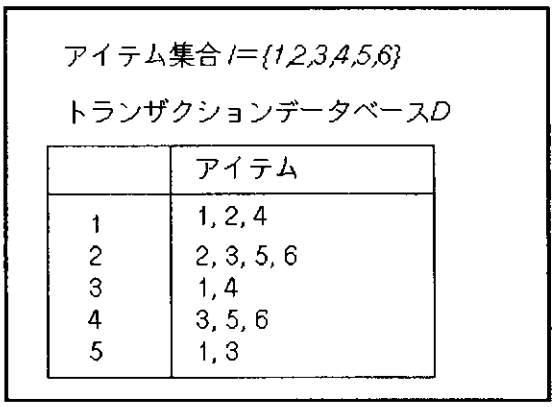

図 1 ：トランザクションデータベースの例

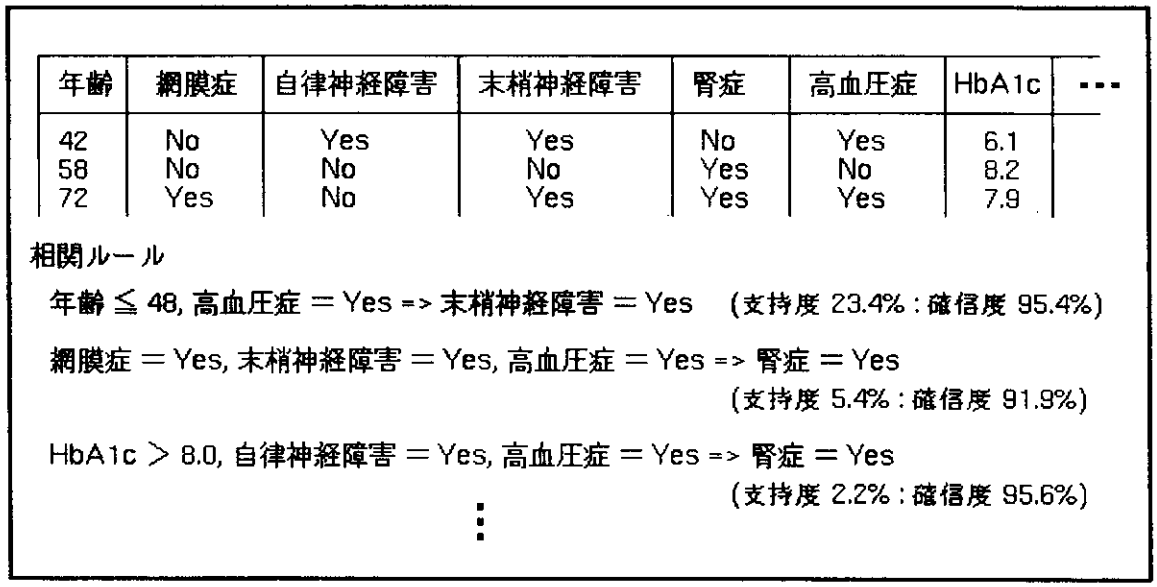

図 2 ：相関ルール発見手法の適用例

1. 結論部に「網膜症」を含むルール

   6528 / 258304 個　　応答時間；29013ms

   例）Disease Duration > 11.5, Family History(Yes), Nephropathy(Yes), Neuropathy-Peripheral(Yes) => Retinopathy(Yes) (Support 5.9% / Confidence 87.1%)

2. 条件部に「高血圧症」を含み、かつ結論部に「HbA1c」を含むルール

   24 / 258304 個　　応答時間：15376ms

   例）Disease Duration ≦ 11.5, Treatment(Diet only), Hypertension(No), Smoking(No) => HbAic ≦ 7.95 (Support 7.3% / Confidence 81.3%)

3. 条件部に「末梢神経障害有」または「自律神経障害有」を含むルール

   3155 / 258304 個　　応答時間：15195ms

   例）Age > 63, Treatment(Insulin), Neuropath-Peripheral(Yes) => Disease Duration > 11.5 (Support 6.2% / Confidence 93.9%)

図３：後解析ツールを用いた解析例とその応答時間

表１：使用した診療データベース

	糖尿病データベース	髄膜炎データベース
収集方法	本大学病院での診療データ	KDD Challenge 2000 にて公開
事例数	1251	140
属性数（前処理前）	60(離散属性 31 / 連続属性 29)	38(離散属性 19 / 連続属性 19)
属性数（前処理後）	22(離散属性 18 / 連続属性 4)	33(離散属性 14 / 連続属性 19)
主な前処理	複数の属性の統合・離散属性値の集約	集約属性の利用

表２：ルール削減手法の適用結果

データセット	最小支持度(%)	適用無	Filter1	削減率(%)	Filter2	削減率(%)	枝刈り	削減率(%)	枝刈り＋Filter1	削減率(%)
糖尿病データベース	10	15001	4995	66.7	8491	43.4	8485	43.4	3830	74.5
	5	19626	7064	64.0	11060	43.6	11036	43.8	5366	72.7
	2	24281	9303	61.7	13845	43.0	13529	44.3	7006	71.1
	1	27514	11174	59.4	16121	41.4	15443	43.9	8457	69.3
髄膜炎データベース	60	904	86	90.5	449	50.3	192	78.8	85	90.6
	50	2896	278	90.4	1341	53.7	602	79.2	229	92.1
	40	5053	485	90.4	2233	55.8	1055	79.1	391	92.3
	30	9150	892	90.3	4111	55.1	1842	79.9	705	92.3

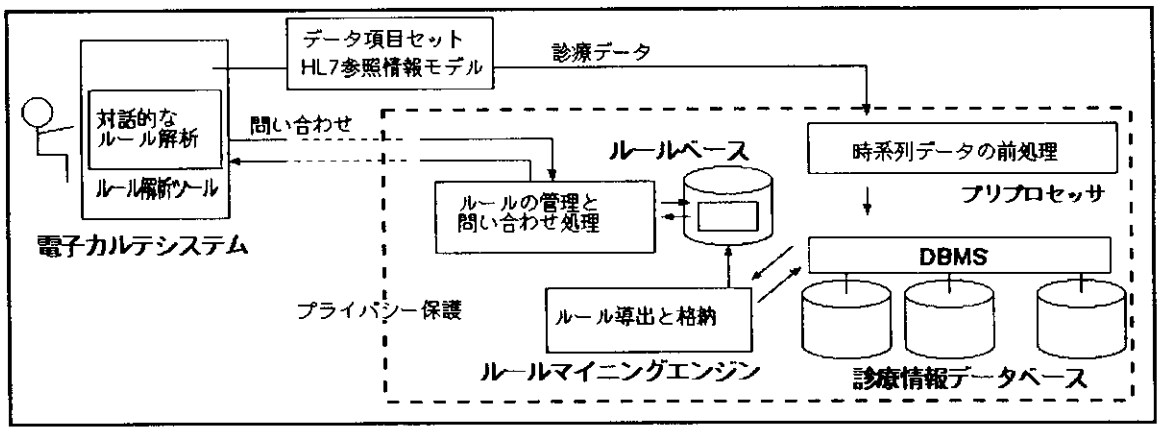

図４：DEBM システムのフレームワーク

標準データ項目セットを用いた知的データベースのセキュリティに関する研究

分担研究者　坂本　憲広　九州大学医学部附属病院 医療情報部 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ら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。
昨年度は、患者プライバシを保護するために、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した。しかしながら、診療データベースには患者基本情報以外にも個人情報が多く含まれる。例えば、診療を行った医師やあるいは家族歴として入手された個人情報などである。
今年度の研究では、患者以外のこうした個人情報のプライバシを保護する手法を提案する。

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．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、昨年度は、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した。
しかしながら、この方法では、診療データベースに含まれる患者以外の個人情報についてはプライバシ保護ができない。診療データベースに含まれる患者以外の個人情報とは例えば、診療を行った医師やあるいは家族歴として入手された個人情報などである。
今年度の研究では、患者以外のこうした個人情報のプライバシを保護する手法を提案する。

昨年度の手法の最大の問題点は、J-MIX に基づき、患者基本情報のみを分離して管理するようにしたことにある。J-MIX は診療情報を交換する上で、非常にコンパクトで有効であるが、一方で、実装や利用者の理解を容易にするため構造が平板(flat)になっており、個人情報である名前や生年月日、住所などが、患者、主治医、紹介医などに分類されて分散して、存在する。例えば、

MD0010050　氏名セット　患者基本情報識別情報 患者.氏名
MD0010370　氏名セット　患者基本情報関係者の連絡先情報　患者関係者.連絡先.氏名
MD0010490　氏名セット　患者基本情報戸籍登録情報　戸籍筆頭者.氏名
MD0010560　氏名セット　患者基本情報世帯登録情報　世帯主.氏名
MD0010670　氏名セット　患者基本情報配偶者情報　配偶者.氏名
MD0010800　健康保険セット 氏名セット 健康保険・福祉情報 健康保険情報　健康保険.被保険者.氏名

などである。
このようにセンシティブな個人情報が分散していては、そのセキュリティを保護することは大変困難になる。
そこで、本年度の研究では、データベース構造を全面的に HL7RIM ベースとすることで、氏名や生年月日などのセンシティブな個人情報を一元的に管理する方法を開発した。

（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．結果

データベース構造設計
昨年度は、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、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のみを HL7RIM にマッピングした。
すなち、HL7 RIM のうち、J-MIX の患者基本情報に関係した部分のみ、すなわち下記のクラスのみをデータベースに実装した。

Entity クラス
Entity_name クラス
Living_subject クラス
Person クラス
Role クラス
Organization クラス
Role_relationship クラス
Employee_Employer クラス

そのマッピング例を表 1 に示す。

表 1

J－M I X	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>

我々はすでに、HL7RIM が J-MIX の上位互換であり、J-MIX で記述される情報は HL7RIM で記述可能なことを示した。そのマッピング例を表 2 に示す。さらに、処方や食事療法、運動療法などの診療情報も HL7RIM で記述可能なことを示した。運動療法に関する診療情報 HL7RIM を用いて記述する方法を図 1 に示す。
上記の研究結果より、HL7RIM 扱うことのできるデータベース構造を設計すれば J-MIX を含む診療情報が格納できることが分かった。そこで本年度は、このデータベースに実装するクラスを HL7RIM 全体に拡張した。
その結果、PostgreSQL データベース上で約 800 の

テーブルを実装した。

行う際にはこれらのテーブルへのアクセスを制御す

表 2

J-MIX	HL7 RIM
健康保険.被保険者.記号	Helthcare_benefit_product_policy.id:SET<II>
健康保険.被保険者.番号	Helthcare_benefit_product_policy.id:SET<II>
健康保険.被保険者.氏名	Entity_name.nm:EN
健康保険.被保険者.姓	Entity_name.nm:EN
健康保険.被保険者.名	Entity_name.nm:EN
健康保険.被保険者.カナ氏名	Entity_name.nm:EN
健康保険.被保険者.カナ姓	Entity_name.nm:EN
健康保険.被保険者.カナ名	Entity_name.nm:EN
健康保険.被保険者.住所	Person.addr
健康保険.被保険者.住所.国コード	Person.addr
健康保険.被保険者.住所.都道府県名	Person.addr
健康保険.被保険者.住所.市区部名	Person.addr
健康保険.被保険者.郵便番号	Person.addr
健康保険.保険者.住所	Organization.addr
健康保険.保険者.住所.国コード	Organization.addr
健康保険.保険者.住所.都道府県名	Organization.addr
健康保険.保険者.住所.市区部名	Organization.addr
健康保険.保険者.郵便番号	Organization.addr
健康保険.保険証.交付年月日	Healthcare_benefit_product_policy.effective_tmr:IVL<TS>
健康保険.保険証.有効期限	Healthcare_benefit_product_policy.effective_tmr:IVL<TS>
健康保険.本人家族区分	Healthcare_benefit_coverage_item.covered_parties_cd:CE
健康保険.入院時負担率	Healthcare_benefit_coverage_item.qty : REAL
健康保険.外来時負担率	Healthcare_benefit_coverage_item.qty : REAL
公費負担.名称	Helthcare_benefit_product_policy.benefit_product_nm:ST
公費負担者.番号	Organization.id:SET<II>
公費受給者.番号	Helthcare_benefit_product_policy.id:SET<II>
公費負担証.番号	Helthcare_benefit_product_policy.id:SET<II>
公費負担証.交付年月日	Healthcare_benefit_product_policy.effective_tmr:IVL<TS>
公費負担証.有効期限	Healthcare_benefit_product_policy.effective_tmr:IVL<TS>
公費負担証.特記事項	Healthcare_benefit_product_policy.benefit_product_desc:ED

ればよい。

この HL7RIM データベースでは、J-MIX では分散していたセンシティブな個人情報がすべて個人を表す Person テーブルや組織を現す Organization テーブルに集約される。そのため、データマイニングを

D．考察

昨年度は、センシティブな個人情報をその他の診療情報から隔離する方針で、そのセキュリティを確保することを行った。しかしながら、この手法では患者基本情報などの情報以外の、例えば医師に関する個人情報などは効果的に保護できないことが判明した。この問題を解決するためには分散するすべての個人情報を洗い出し、それらを分離し管理する必要がある。このような処理を行うとデータマイニング上重要な情報が失われる可能性がある。例えば、ある医療従事者を介した院内感染などがその例である。本研究は、日々刻々集積される診療情報を基にして、こうした院内感染などの傾向も早期に指摘できることを目的としており、こうした情報が失われてしまうデメリットは非常に大きい。

一方、本年度行った拡張により、全ての個人情報を一元的に管理することができるようになった。このことにより、データマイング時に必要な情報をすべて利用できるとともに、プライバシー保護の観点からは、Person テーブルや Organization テーブルなど特定の少数のテーブルのみを監視し、その情報が漏洩しないようにすればよいため、効率的なセキュリティ管理が実現できるものと期待される。

しかしながら、センシティブな情報がその他の情報と同じ管理レベルに存在することについては、センシティブな情報を含むテーブルのみ暗号化するなどの新たな方策を考案する必要がある。

また、テーブルごとや情報内容によるアクセス権の設定方法も今後詳細な研究を必要としている。

E．結論

センシティブな個人情報を扱う際には、患者基本情報と診療情報の分離する方法と、患者、医師など荷かかわらず、個人情報を一元的に管理する方法がある。本研究では、両者の方法について、システム面および情報構造の観点から、そのメリットとデメリットを分析し、されには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。

一方、医師や臓器提供のドナーの情報も診療情報には含まれる。これらの情報は、院内感染や血液感染などのデータマイニングには必要な情報である。したがってこれらの情報をデータマイニングに利用可