医療情報システムのデータマイニングによる関連病名の発見*

ジャンミシー パッタモン[1] 金谷 敦志[1] 梅村 恭司[1] 古田 輝孝[2] 櫻井 潤児[2] 木村 通男[3]
豊橋技術科学大学 情報工学系[1] 株式会社ＮＴＴデータ東海[2] 浜松医科大学[3]

1 はじめに

医療情報システムにおいて「同じ病気の患者のデータを集めたい」という要求が高い．しかしながら，病名については，同類の病名でありながら，微細な違いにより，別の病名を医師が付与するため，異なる病名とされることが多い．また，ほぼ同一の病名であっても，医師の個人差により，病名に対する表記のゆれが存在する．このため，同じ病気の患者のデータを集めることは簡単な操作ではない．

この問題を解決するために，本研究は表記のゆれがある病名と，それに対する検査情報・処方の情報の組をもとに，「同じような検査結果であり，かつ，同じような処方を行う」という観点から病名の組を構築する実験を行う．それによって，病名の表記ゆれに対処できる情報が得られるかを調べた．

2 関連病名の判定のための確率

二つの病気 A と B に関連があるならば，病気に関係する検査 i を前提とした確率 P(A|i)と P(B|i)から関連度が計算できると考えた．

2.1 スコア関数

病気の事象を A，処方の事象あるいは異常と判定された検査を i と考えると，病気 A に対して処方 i が処方される確率，又は異常と判定された検査 i が起こる確率はは P(i|A)と記述する．この時，２つの病名の関係を求めるために次の関数を用いる．

$$score\ (A,B)=\frac{\sum_{i\in V}\log\ \min\{\frac{P(i\mid A)}{P(i)},\frac{P(i\mid B)}{P(i)}\}}{i\text{の数}} \qquad (1)$$

但し，A，B は病名，V は A または B に対する処方、又は、異常と判定された検査，$\frac{P(i\mid A)}{P(i)}$ は処方又は異常と判定された検査 i が独立して起こる確率と病気 A に対して処方される又は異常と判定された検査の起こる確率の比である．

2.2 処方と検査結果による統計検定

処方と検査結果による２つの病名の関係を求めるために，上記でそれぞれ求めた処方による統計検定のスコアと検査結果による統計検定のスコアの和を取る．

$$addScore(A,B)=score_{\text{処方}}(A,B)+score_{\text{検査結果}}(A,B) \qquad (2)$$

2.3 出力の評価

スコア計算の結果はどれくらいの関連があるかは専門家の医師が行い，次の 4 段階で２つの病名の関係性の評価点数を付ける．

同じと思える	：４点
包含関係	：３点
関連あり（一つの症状である）	：2 点
関係なし	：1 点

3 病名に対応する処方情報と検査情報の前処理

患者名と患者の個人情報を削除し，病名に対応する処方情報と検査情報のデータの前処理を行う．研究に用いた医療情報データは医療環境における電子データ交換用“Health Level 7(HL7)”といった医療情報標準化規格に準したデータである．浜松医科大学において，過去８年間の処方，検査結果，病名登録の情報は，NEC 社製の病院情報システムから，HL7 形式で，NTT-Data 製の臨床情報検索用データベースに送られており，このデータの形式は各患者に対する病名，処方，検査結果の情報が別のファイルに保存され，それぞれのファイルに日付が記録されている．このため，データの前処理として，同じ患者 ID かつ同じ日付が記録されているデータに対して，病気の名前とそれに対する処方コード，および，病気の名前と検査結果が異常と判定された検査のコードのリストを作成する前処理を行った．

また，痛み止め薬などの病気と無関係に処方されるような処方は情報量が低く，スコア関数の計算のノイズになる可能性があると考えられるので，これらの処方を計算対象にしないことが望ましい．そのため，先に各処方の出現確率を計算して，昇順にソートする．昇順にソートした処方の出現確率のリストを小さい順から 2 割だけの処方を残して，対象する処方のリストを作成する．それ以上出現する処方は情報量が低いと仮定し，対象外の処方とする．その後，病気の名前と対象する処方リストにあるその病気に対する処方のリストを作成する．2 割という数字は，処方内容と病名との関連があるかという医師の判断に従ったものである．

4 統計検定実験

4.1 入力データ

HL7 医療情報データ 2000 年 1 月（80MB）

4.2 実験項目

(a) 病気に対する全ての処方によるスコア関数の計算
(b) 処方の処置された確率を昇順にソートして小さい順から 2 割以上をカットした処方によるスコア計算
(c) 異常と判定された検査によるスコア関数の計算
(d) (b)と(c)の結果による addScore 関数の計算

*Finding related disease names by data mining of a medical information system

[1] Toyohashi University of Technology Dept. of Information and Computer Sciences
[2] NTT DATA TOKAI Corporation
[3] Hamamatsu University School of Medicine

4.3 実験結果と評価

スコアの高い順から 5 位までの出力例とその評価を表 1 に示す.

表 1(a)：処方によるスコア計算の出力例と評価

順位	スコア	病名	病名	評価
1	7.3908	トリコモナス感染症	トリコモナス症	4
2	7.1880	全顎的萌出遅延	左）頸部リンパ節炎	1
3	7.1880	（上顎）義歯不適合	緑内障	1
4	7.0691	ポルフィリンン症の疑い	脱毛症	1
5	7.0084	卵巣悪性腫瘍の疑い	慢性副鼻腔炎の疑い	1

表 1(b)：ノイズを減らした処方によるスコア計算の出力例と評価

順位	スコア	病名	病名	評価
1	5.5373	地図状舌	舌痛症	2
2	5.5373	両）乱視	眼精疲労	3
3	5.5373	両）乱視	（両）水晶体亜脱臼	2
4	5.5373	更年期障害	血栓症の疑い	2
5	5.5373	更年期障害	肝機能低下の疑い	2

表 1(c)：検査結果によるスコア計算の出力例と評価

順位	スコア	病名	病名	評価
1	3.0204	先天性副腎過形成	末梢神経障害	2
2	3.0204	思春期早発症	末梢神経障害	1
3	3.0204	左母趾関節炎	末梢神経障害	1
4	3.0204	筋肉痛の疑い	末梢神経障害	2
5	3.0204	ふらつき	末梢神経障害	3

表 1(d)：ノイズを減らした処方と検査結果による addScore の出力例と評価

順位	スコア	病名	病名	評価
1	8.4806	更年期障害	血栓症の疑い	2
2	8.4806	更年期障害	肝機能低下の疑い	3
3	8.4806	血栓症の疑い	肝機能低下の疑い	2
4	7.0943	更年期障害	卵巣機能不全	3
5	7.0943	血栓症の疑い	卵巣機能不全	2

4.4 考察

全ての処方によるスコア関数の計算結果は，ノイズが大きく，関係ありそうもない病名のスコアが高くなってしまった．これに対して，処方が処置された確率を昇順して小さい順から 2 割以上の処方をカットして，ノイズを収めることができた．得られた結果は全く同じ病名が求められるまででではないが，ある病気とその病気の一つの症状として表す病気という関連を持つ病名の組が得られたと言える．一方，検査結果が異常と判定された検査による実験は目的に合致した組が得られなかった．これの影響で，処方と検査による計算の和と取った addScore 関数の結果も目的に合致していないと考えられる．

5 処方による実験の拡張

前述した結果より，処方された確率を昇順して小さい順から 2 割以上の処方をカットして計算した結果は他の結果よりもっともいい評価が得られるとわかったので，処方の処置された確率を昇順して小さい順から 2 割以上の処方をカットする方法を用いて，入力するデータサイズを増やす．さらに，前処理の段階で，患者の入院外来区分を入院と外来に分けて実験を行ってみた．

5.1 入力データ

HL7 医療情報データ 2003 年 8 月～11 月（382MB）

5.2 実験結果と評価

スコアの高い順から 5 位までの出力例を表 2 に示す.

表 2(a)：外来入院に区分されていない時の出力例とその評価

順位	スコア	病名	病名	評価
1	7.0884	右）環指挫傷	環指挫傷	3
2	6.7418	皮膚悪性腫瘍	有棘細胞癌	3
3	6.3952	乳癌の疑い	末端肥大症の疑い	1
4	6.3952	椎骨脳底動脈循環不全	右）慢性中耳炎急性増悪	1
5	6.3952	椎骨脳底動脈循環不全	右）顔面神経麻痺	1

表 2(b)：外来区分の出力例とその評価

順位	スコア	病名	病名	評価
1	7.4383	膝関節周囲炎	股関節炎	3
2	7.4383	乳癌の疑い	末端肥大症の疑い	1
3	7.4383	上気道炎（妊娠３２週）	腹式帝王切開術	1
4	7.4383	重症筋無力症の疑い	喉頭炎の疑い	1
5	7.4383	歯周炎	歯肉腫瘍（良性）	2

表 2(c)：入院区分の出力例とその評価

順位	スコア	病名	病名	評価
1	5.4071	腟造設術後	腟炎	2
2	5.4071	抑うつ状態	糖尿病：〈DM〉の疑い	1
3	5.4071	抑うつ状態	高脂血症：〔その他及び詳細不明〕の疑い	1
4	5.4071	不整脈の疑い	抑うつ状態	1
5	5.4071	不整脈の疑い	脳腫瘍の疑い	1

5.3 考察

外来入院に区分されていない処方によるスコア計算出力，外来区分の処方によるスコア計算出力，入院区分の処方によるスコア計算出力を比較すると，入院区分によるスコア出力と，外来区分のスコア出力は目的に合致した組が得られなかった．関連性がある病名のペアを求めるためには，外来入院に区分されない処方によるスコア計算が必要であることがわかった．これはそれぞれ外来区分の病名と入院区分の病名の間に有効な関連名があることを示唆している．

6 今後の課題

ノイズを減らす方法として 2 割の他にいろいろな割合で試して，効率的に処方によるノイズを削除する方法を検討することが今後の課題である．また，今回行った実験は病名のシソーラス作成に向けての初段階であって，簡単な検定モデルを利用したが，統計検定モデルをさらに改善し，医療情報を利用してより正確的に関連性の高い病名の組を求めることを今後の課題として考えている．

外来入院に区分してはいけないという実験結果については，予想に反する結果であり，更に詳しく調べたい．

7 参考文献

[1] 木村　通男．医療情報標準化規格．HL7 医療情報標準化規格-その概略-．医療科学社，2002

[2] 河野　崇，古田　輝孝，村井　靖ら．オーダエントリシステムから HL7 を介してデータを集積する，柔軟迅速な検索を可能としたデータウェアハウス．第 22 回医療情報学連合大会論集．医療情報学　2002；22：761-762

[3] ICD-10 対応電子カルテ用標準病名マスターVer2.1. 財団法人医療情報システム開発センター，2002

[4] JAHIS 臨床検査データ交換規約＜オンライン版＞ ver.1.0. 財団法人医療情報システム開発センター,1999