様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　　６　日現在

機関番号：　３４５０４
研究種目：　基盤研究(A)（一般）
研究期間：　2012～2015
課題番号：　２４２４００２５

研究課題名（和文）医薬副作用に特徴的な化学構造マイニングと早期シグナル検出

研究課題名（英文）Chemical structure mining for adverse reactions and early stage signal detection

研究代表者
岡田　孝（Okada, Takashi）

関西学院大学・理工学部・理工学部研究員

研究者番号：００１０３１３５

交付決定額（研究期間全体）：（直接経費）　　28,800,000円

研究成果の概要（和文）：医薬品の有効成分をもとに副作用解析を行うためのPharmCompoデータベースを開発した．塩の除去やプロドラッグの変換を行うとともに，ATC分類コードを付与している．ATC分類樹の中で，特徴的に副作用の多発する分類ノードや成分ノードを検出するアルゴリズムを開発した．JADER副作用報告データベースに対し，アナフィラキシーショックなど7種の副作用についてこのアルゴリズムを適用し，副作用が多発する特徴的な医薬品分類や，個別の医薬品を検出した．さらにそれらノード群での構造の視察による種構造の発見および構造精錬により，それぞれの副作用に対する構造アラートを提案することができた．

研究成果の概要（英文）：Effective components of drugs are accumulated in PharmCompo database. Each entry in the database has ATC codes. A new algorithm has been proposed to detect classification nodes and component nodes characteristically related to an adverse reaction by drugs. We have applied this algorithm to adverse event reports in JADER using anaphylaxis and other 6 reactions. Drug classifications and component drugs were detected causing frequent reactions. Browsing the structures of these drugs led to several substructures, and the succeeding structure refinement process enabled the proposal of structural alerts to these adverse reactions.

研究分野：　データマイニング

キーワード：　医薬品副作用　データマイニング　ATCコード　分類樹説明変数　構造アラート

１．研究開始当初の背景

(1) 新規医薬品による特異体質毒性では，上市後の副作用報告からの早期発見が重要である．わが国では PMDA が，重篤副作用症例を集積し，ROR 下限値を用いたシグナル検出法を個別医薬品のスクリーニング作業に用いている．

(2) 上記の方法は個別医薬品の副作用報告数をすべての医薬品の報告数と比較するものである．副作用も一種の生理活性である以上，副作用発現に特徴的な化学構造が存在するはずである．そのような部分構造を発見できれば，医薬品群に対するシグナル検出が可能となる．

(3) 具体的には，特定の構造が特定の副作用を起こしやすいことが分かれば，あらかじめ当該副作用への注意を喚起することができる．また同一部分構造を有する医薬品群中で，特定の副作用を起こしやすいかを検出できる．

２．研究の目的

重篤副作用症例のデータベース JADER から，代表的な副作用を対象として，それぞれの副作用を起こしやすい基本活性構造の BAS 群を抽出する．その有効性を米国のデータベース FAERS を用いて検証する．得られた BAS 群は BASiC 知識ベースにおいて公開する．対象とする副作用は，肝障害，アナフィラキシーショック，横紋筋融解症など１０種程度とする．また，活性構造抽出のためのグラフカーネルを利用した新たな高速マイニング技法を開発する．

３．研究の方法

(1) マイニングを遂行するための基盤として，副作用と化学構造，医薬品名および医薬品分類を整理したデータベース PharmCompo を構築する．対象とする元のデータベースは JADER, FAERS, SIDER2, DrugBank, および日本の医薬品添付文書とし，有効成分の化学構造を基本単位として整理を行う．医薬品分類としては ATC コードを採用する．

(2) PharmCompo 中の全医薬品を対象として，特定の副作用群について，特徴的な活性構造を抽出する．

(3) ATC 分類樹を説明変数と見なし，特定の副作用を起こしやすい分類樹ノードを検出するためのアルゴリズムを開発する．

(4) 検出された分類樹ノードが中間ノードの場合，その分類下の医薬品群に特徴的な部分構造を視察により検出する．

(5) 検出された分類樹ノードが葉ノードの場合，他の葉ノード医薬品構造との比較から，当該医薬品が副作用を起こす原因と考えられる部分構造を推定する．

(6) 上記作業により副作用毎に構造アラートを検出する．

(7) グラフカーネルを用いた新規グラフマイニング法を開発する．

４．研究成果

(1) PharmCompo データベースの開発

副作用報告では実質的に同一の医薬品であっても，酢酸塩とシュウ酸塩のような微細な差により異なった医薬品として扱われる．そこで医薬品の有効成分の化学構造に着目したデータベースを開発した．これは下記 5 種の医薬品データベースに出現するすべての低分子有機化合物の有効成分を収載したものである．①JAPIC2012, ②DrugBank V3.0（approved と withdrawn），③SIDER2, ④ JADER 2003.4Q~2013.1Q, ⑤ FAERS 2010Q1~Q4(JAPIC AERS より収録).

基本となる有効成分テーブルには，校訂済みの構造式を SMILES 表記で記述しているが，その際に次の処理を施している．①プロドラッグは代謝後の構造に置き換え，②Tautomer の構造記述を標準化，③Markush 式は代表構造を収録，④塩は水素原子に置き換え，⑤光学活性情報は削除．

さらにこれら有効成分とソース DB 所載医薬品との対応テーブルを整備するとともに，一般利用者が容易に使えるハイパーリンクを付与した Excel 表も利用可能とした．現在の PharmCompo データベースは，１万件近いソース DB の医薬品名称を，2460 種の有効成分で整理したものとなっている．これらのデータを整理したことにより，JADER の報告数を有効成分毎に正確に数え上げることが可能となった．

(2) ATC 分類コード

PharmCompo データベースの各成分には，WHO Oslo center が管理公表している ATC コード(Anatomical Therapeutic Chemical Classification System) も付与した．例えば鎮痛剤の aspirin に与えられている N02BA01 のコードは次の意味を持つ．①A レベル： alphabet N — 神経系に作用，②T レベル： 数字 02 — 鎮痛薬，③P レベル：alphabet B — 「その他の鎮痛剤および解熱薬」サブグループ，④C レベル：alphabet A — サリチル酸またはその誘導体．なお，最後の数字 01 はこの分類内での個々の医薬品に与えられた連番である．aspirin は抗血栓剤としても処方されており，そのために別コード B01AC06 も付与されている．ATC システムには現時点で 1000 種以上の分類ノードが設定されている．

PharmCompo 中の，1876 種の有効成分には WHO で与えられている 2376 コードを利用した．しかし，584 種のコード未付与の化合物が存在した．これらについては，添付文書を調査し，ATPC レベルまでの 5 桁のコード 657 種を割り当てた．これにより，ATC

コードを利用した解析が全ての低分子有機化合物医薬品を対象として行える．

(3) 特徴的分類ノードの抽出アルゴリズム

ATC コードは医薬品の樹状分類であり，図書の 10 進分類や生物分類と同様の性質を有していると考えられる．ただし，葉ノードの個別医薬品は複数の C レベルノードからリンクされている場合がある．

決定木や回帰木のように属性の値により分類を創りだそうとする試みは，機械学習で長い歴史を有する．しかし，２値分類の問題を扱うときに，このような分類樹が説明変数として与えられており，その中から特徴的なノードを選択する試みは余り行われていない．本研究では，ATC 分類体系から，特定の副作用に特徴的なノードを選択する課題を設定し，そのアルゴリズムを提案した．

まず，下の図 1 に示すように分類樹中で特徴的か否かの判定対象となるノードを node，その親ノードを parent，i 番目の子ノードを $child_i$ とする．また，分類樹全体の根となるノードを root とする．

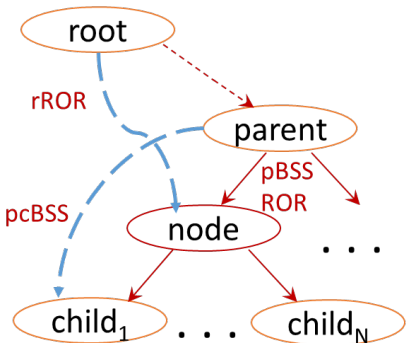


図 1. 分類樹のトポロジー

ここで我々が選択したいノードは，parent 下のノード群中で明確に毒性の強いものであろう．その判定基準には，シグナル検出での ROR の 95%信頼限界下限値 $lowROR$ を採用することが適当であり，次の条件 1 を課す．ここで ROR の計算では，parent 下の全医薬品中での node 部分の ROR を計算する．λとしては 1.0 を採用する．

【条件 1】$plowROR > \lambda$

この node が兄弟ノード群に比して高毒性を示していたとしても，それが全医薬品の平均的な値よりも低ければ，興味あるノードとはいえない．また，余りに事例数が少ない場合は，ノイズである可能性が高い．そこで，次の条件 2 を導入する．ここで，$rlowROR$ は，データベース中のすべてを対象とした時の $lowROR$ である．μとしては 5.0 を採用する．

【条件 2】$rlowROR > \mu$ かつ事例数 > 3

もし上記の 2 条件を満たす node があっても，child ノード群中での少数の高毒性ノードに引きずられた結果である場合，我々が注目すべきは node ではなくこれらの少数 child ノードであろう．parent から node への移動による識別力の上昇が，parent からこれら少数 child ノードへの移動による識別力の上昇よりも大きい場合にのみ，node を特徴的ノードとして採用すべきである．

識別力としては，名義変数の平方和分解により定義される群間平方和 BSS を使うのが適当である．n 件の事例群中で値αを持つ事例の確率を $p(\alpha)$，これらを g 個の群に分割した時，各群での値αの確率を $p_g(\alpha)$とすると，以下の式で群 g への群間平方和 BSS^g が定義できる．

$$BSS^g = \frac{n^g}{2}\sum_{\alpha}\left(p^g(\alpha) - p(\alpha)\right)^2$$

この識別力 BSS を利用して，次の条件 3 を導入する．ここで $pBSS$ は parent と node 間の BSS を表し，$pcBSS_i$ は parent と $child_i$ 間に仮想の親子関係を想定した場合の BSS である．また，合計は条件 1, 2 を満たす child ノード群についてのみ計算する．

【条件 3】$pBSS > \sum_i^{toxic} pcBSS_i$

これまでの 3 条件を満たす node で，子ノードが 1 個しかない場合が現れる．このような場合，この node の代わりに child を特徴的ノードとして選択すべきである．そこで以下の条件を導入する．

【条件 4】node 下に単一の child node しかない場合は，それを注目ノードとする．

(4) 副作用への適用

前節の特徴的分類ノード抽出法を，ATC コードを説明変数として JADER データベースに適用した．対象副作用をアナフィラキシーショックに設定した結果を示す．なお JADER の疾患名に"アナフィラキシーショック", "アナフィラキシーショック（N）", "アナフィラキシー性輸血反応", "アナフィラキシー反応", "アナフィラキシー様ショック", "アナフィラキシー様反応"のいずれかを含むものとした．また，対象医薬品は第１被疑薬でかつ単成分のものに限った．

すべての ATC 分類ノード（個別成分を含む）に対する，$plowROR$ vs $rlowROR$ および $pROR$ vs $rROR$ の対数軸での散布図を，次の図 2 と図 3 に示す．条件 1，2 として $plowROR$>1.0，$rlowROR$>5.0 を採用したので，これらの図での赤点が抽出した結果となる．なお右上領域で，条件 3 を満たさなかった点は紫色に，事例数が 3 に満たなかった点は薄青色に表示されている．この散布図から，提案法は外れ値に該当する特徴的な分類ノードを拾い上げていることが分かる．

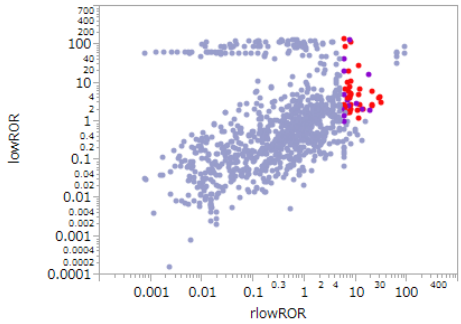


図 2.　ATC コードの *lowROR* 値の散布図

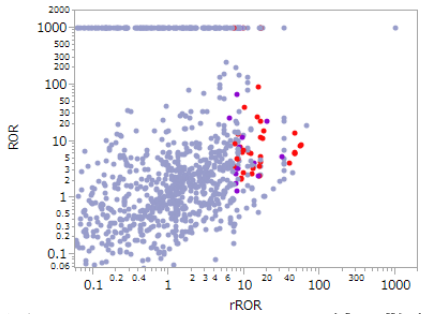


図 3.　ATC コードの *ROR* 値の散布図

選択した特徴的ノードは 43 個であり，その内訳は ATP レベル:1 種，ATPC レベル: 7 種，個別成分レベル: 35 種 23 成分であった．分類ノードの選択結果を表 1 に，個別成分レベルの選択結果の内から 10 種を表 2 に示す．個別成分は局所麻酔薬のように複数の葉ノードとして出現するものがある．表 2 の右端の列には，1 回のみ現れる場合はそのコードを，複数回現れる場合は ATPC レベルの出現数を示した．

表 1．特徴的な分類ノード

ATC code	Percent	*rlow ROR*	Name
R02A	18.1	6.1	THROAT PREPARATIONS
A03AX	19.5	6.1	Other drugs for functional gastrointestinal disorders
A04AA	35.0	11.9	Serotonin (5HT3) antagonists
B05CA	30.8	7.5	Antiinfectives
C05AD	20.0	6.7	Local anesthetics
J01DB	32.2	13.7	First-generation cephalosporins
S01HA	20.7	7.0	Local anesthetics
S02DA	20.0	6.7	Analgesics and anesthetics

表 1 の特徴的なノードに予見力があるか否かは現在調査中である．しかし，アナフィラキシーショックは体質的な要因が大きいと考えられてきた疾患であり，このような特徴的分類が存在することはこれまで考えられていなかった．

表 2．特徴的な成分ノード

Drug name	percent	*rlowROR*	ATC code
Bromhexine	27.0	6.0	R05CB
Butylscopolamine	31.9	11.5	A03BB
Cefaclor	54.3	31.5	J01DC
Cefmenoxime	62.5	29.2	J01DD
Cefoperazone	63.3	27.3	J01DD
Cefpirome	22.7	6.7	J01DE
Ceftriaxone	27.9	11.7	J01DD
Chlorhexidine	33.3	8.1	7
Dextromethorphan	22.7	6.3	R05DA
Dibucaine	58.3	20.7	5

これらの結果からいくつかの知見が得られる．例えば，抗てんかん薬はアナフィラキシーの要注意医薬とされているが，特徴的に高い分類としては現れない．

R02A, A03AX, S01HA, S02DA の様に，身体が外部との接触する部位に係る医薬品群が多い．個別成分でも bromhexine, pranoprofen, butylscopolamine, dextromethorphan は身体外部との接触部位に作用する．免疫システムがこれらの組織で発達しているためアナフィラキシーが起こりやすいと想定される．

また，抗生物質が多く現れるので，tree 構造で結果を整理したところ，第 1 世代から第 4 世代セファロスポリンとカルバペネムでは，この順に発症頻度が低下していることが分かる．

(5) 構造アラートの発掘

特徴的な分類ノードや成分ノードの化合物構造を視察すれば，それらの要因と考えうる部分構造を容易に認識することができる．その構造を種として入力し，データベース全体を参照しながら，より適切な部分構造へと構造精錬するシステムを開発した．これを利用して，得られたアナフィラキシーに対する構造アラートを次表に示す．この表で，T/N の欄は，当該部分構造を有する医薬品でアナフィラキシーの発生頻度が 5%以上のものを T，それ以下のものを N として，それらの比を示したものである．なお，全医薬品ではこの比は T/N=183/877 である．

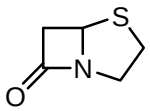
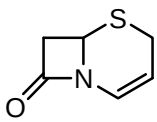
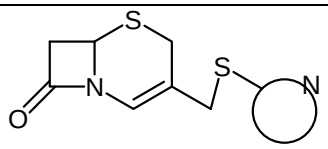
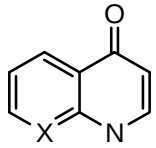
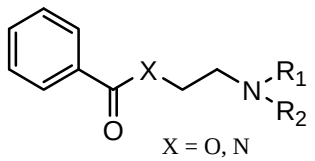
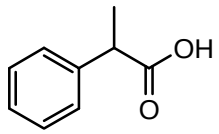
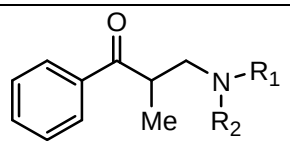
この表で，1 行目は penicillin, 2 行目は cepharosporin, 3 行目は quinolone である．これらの部分構造はすでによく知られたものである．しかし，2'の構造は 2 から構造精錬で得られた構造であり，本研究で初めて明らかになったものといえる．チオエーテル構造を介して含窒素複素環が置換基に存在する場合は，11/0 と全てが高いアナフィラキシー頻度を示している．

このような解析により，アナフィラキシーに特徴的な構造アラートを指摘することができた．

このような解析を，アナフィラキシー以外にも，間質性肺疾患，肝機能異常，血小板減

少，好中球減少，貧血および横紋筋融解症について，JADER DB の解析を行い，特徴的なノードと構造アラートを検出することができた．詳細は投稿予定の論文に譲る．

表 3．アナフィラキシー構造アラート

	Structure	T/N
1		7/5
2		25/6
2’		11/0
3	 X = C, N	11/6
4	 X = O, N	6/4
5		10/2
6		3/0

(6) 分子グラフとクラスの関係性を学習するには，2 つのグラフの類似度を正確に測る指標やそのアルゴリズムが重要となるが，この問題に対して，グラフの類似度を測るための指標とアルゴリズムを 2 つ提案した．

① 1 つ目の提案手法は，グラフの各頂点をアダマール符号でラベル付けし，2 つのグラフの各頂点のラベルを比較することで 2 つのグラフの類似度を測る．ただし，頂点同士の比較では，部分構造の類似度を測れないので，ある頂点のラベルとその頂点に隣接する頂点のラベルの和を繰り返しとることにより，その頂点の周辺の部分構造の表現を可能にした．アダマール符号を用いることで，ラベルの和が平均 0 の 2 項分布に従うので，繰り返し和をとったとしても省メモリで動作することを可能にした．

② 2 つ目の提案手法は，1 つ目の改良手法である．まず頂点同士の比較において，一致/不一致で判定していたものを類似の程度に応じて数値化した．また，すべての頂点の組み合わせをとるのではなく，最小 2 部グラフマッチング問題を取り入れることで，最適な頂点の対応関係をとるようにした．これにより 2 つのグラフの類似度をより正確に測ることを可能にした．ベンチマークデータを用いて提案手法の性能を評価したところ，この分野の代表的な従来手法に比べ分類精度が 10% も大幅に向上したケースもあった．また，提案されたアルゴリズムは，グラフの頂点数に対して，多項式時間で動作するので，大規模なデータにも適用できる．

5．主な発表論文等

〔雑誌論文〕（計 7 件）

① 岡田 孝，人工知能 知識発見そしてデータマイニング 30 年の歩みから，J. Comp. Aided Chem.，査読有，Vol.18，2017，印刷中.

② 長村 佳歩，奥井 颯平，猪口 明博, 滑らかな変化を検出するためのグラフ系列クラスタリング，情報処理学会論文誌，査読有，Vol.58, pp. 278 - 287, 2017.

③ T. Kataoka, A. Inokuchi, Hadamard Code Graph Kernels for Classifying Graphs, Proc. of the fifth International Conference on Pattern Recognition Applications and Methods, 査読有 pp.21, 2016.

④ 片岡 哲也，猪口 明博，アダマール符号を用いたグラフカーネルによるグラフクラス分類, 情報処理学会論文誌, 査読有，Vol.57, pp2122 - 2130, 2016.

⑤ N. Takada, N. Ohmori, T. Okada, Mining Basic Active Structures from a Large-scale Database, J. Cheminformatics, 査読有，Vol.5, 1-8, 2013, DOI: 10.1186/1758-2946-5-15.

⑥ 岡田 孝，大森 紀人，ベイジアンネットによる毒性評価システム ToxBay，J. Toxicol. Sci.，査読有，Vol.37 Supplement1, 122, 2012.

⑦ A. Inokuchi, H. Ikuta, T. Washio, Efficient Graph Sequence Mining using Reverse Search, IEICE Transactions, 査読有，Vol.95-D, 1947, 2012, DOI: 10.1587/transinf.E95.D.1947.

〔学会発表〕（計 3 件）

① 岡田 孝, 大森紀人, 堀川祐志, 猪口明博, 分類樹説明変数による副作用自発報告データベースのマイニング，人工知能学会知識ベースシステム研究会，2015 年 8 月 7 日，関西学院大学大阪梅田キャンパス（大阪府・大阪市）.

② 大森紀人，堀川祐志，岡田 孝，横紋筋融解作用への化学構造の影響，第 41 回日本毒性学会学術年会，2014 年 7 月 3 日，神戸コン

ベンションセンター（兵庫県・神戸市）.
③ 大森紀人，堀川祐志，岡田 孝，医薬品統合データベースの作成と ATC コードによる横紋筋融解症の解析，第 42 回構造活性相関シンポジウム，2014 年 11 月 13 日~14 日，くまもと森都心プラザ（熊本県・熊本市）.

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計　0 件）
名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0 件）
名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
BASiC (Basic Active Structures in Chemicals)
http://www.dm-lab.info/BASiC

６．研究組織
(1)研究代表者
岡田　孝（OKADA, Takashi）
関西学院大学・理工学部・理工学部研究員
研究者番号：00103135

(2)研究分担者
猪口　明博（INOKUCHI, Akihiro）
関西学院大学・理工学部・准教授
研究者番号： 70452456

鷲尾　隆（WASHIO, Takashi）
大阪大学・産業科学研究所・教授
研究者番号：　00192815
平成 24 年度のみ

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）