

レセプトデータへのデータマイニング手法の適用による 医薬品副作用発現の予測・予防に関する知識発見

塚本 哲治

I はじめに

近年のコンピュータ技術の飛躍的な進歩により大量のデータの蓄積が容易になったことを背景に、データマイニングを適用してビジネスに役立つ情報を取り出そうという試みが広がっている^{[2] [6]}。例を挙げれば Web サイトのアクセスログ解析、顧客関係管理などのマーケット指向の分析だけでなく、創業支援やゲノムデータの分析、薬剤疫学といった医療分野など広範囲で応用されている^[13]。

このような中、我々が病気に罹ったときに受けた実地医療の情報（データ）は、医療の改善や有害事象の予防のためにどれだけ活用されているのであろうか。現状の医療現場での患者の受ける恩恵を高めるためには、実地医療のデータに潜む重要な情報を抽出し、患者の特性に合わせて治療方法を選択し薬剤の有害事象を予防することが必要と考えられるのである。

現在、医療用医薬品として市販されている薬剤は、臨床試験が実施され有効性と安全性が確認され承認を受けたものである。その有効性・安全性の情報は医薬品添付文書にまとめられている。

医師、薬剤師等の医療関係者は医薬品添付文書の情報に基づいて、医薬品のより安全な使用に砕心している。しかしながら多様な合併症を持つ患者に多種多様な薬剤の逐次併用や同時併用があり、臨床試験では把握できなかった予期せぬ有害事象発現の可能性を孕んでいる。すなわち、実地医療における有害事象の発現は、時として思いがけないものであり、添付文書の情報通りでないことが多い。

米国においては、市販されている薬剤の中毒反応で、毎年 10 万人近い患者が亡くなり、さらに 200 万人が入院するという統計があり^[12]、わが国においても平成 18 年度厚生労働白書において市販後の安全対策の重要性が謳われている^{[9] [11]}。

本研究においては、治療に伴う重篤な有害事象の発生に至ってしまうような薬剤の併用方法や合併症等の存在を明らかにし、より安全な薬剤使用につながる知識を発見することを目的とした。

具体的には、診療報酬明細書（以下、レセプト）データを分析材料として、間質性肺炎

という重篤な有害事象発現と因果関係がある可能性のある薬剤処方や合併症に関するシグナルを検出し、どのような特徴・法則があるのかを発見する。

シグナル検出手法に基づく先行研究として、医薬品副作用情報のデータを分析材料に用いている研究^[5]とレセプトデータを用いている研究^[15]がある。いずれも注目すべき医薬品と注目すべき副作用のシグナルを検出し、医薬品と副作用に因果関係がある可能性を探り出すものである。

またレセプトデータにデータマイニング技術を適用し糖尿病関連の合併症発症の予測モデルの構築を試みている研究がある^[10]。

本研究の新規性は、データマイニング手法の1つである Emerging Pattern^{[3][4]}に基づくシグナル探索手法を新たに開発し、その手法を適用することでレセプトデータに含まれる多種類の情報を活用し、併用薬剤の相互作用、患者の持つ合併症、受けている治療や処置の内容など有害事象に至る原因事象を、より多元的に捉えることができることである。

II 分析材料

医薬品副作用情報へのシグナル分析の適用は、すでに欧米の規制当局で実施されており、日本においても導入検討の段階である^[5]。これらの分析材料は自発報告をベースとして集積された副作用情報であるが、自発的な報告であることから医師や医療機関の認識の違いで報告されないケースがあることやメディアからの情報によるバイアスで報告頻度が増減するという問題が含まれている。このような自発報告データの弱点を避けるべく、本研究においては治療内容および薬剤処方に関する生の情報を含んでいる有益な情報源であるレセプトデータの利用を試みた。

レセプトとは医療機関が患者1人の1ヵ月分の治療、検査や調剤に要した費用を保険点数としてまとめ、保険支払い者（健康保険組合や市町村）に対して医療費を請求するための明細書であり、月毎に発生する傷病、薬剤、処置等の多様なイベントを含む。これらの各イベントをアイテムとよぶ。

本研究で用いたレセプトデータは、健康保険組合等と契約し医科ならびに調剤のレセプトをデータベース化して運用している^[14]日本医療データセンター社（以下JMDC）と関西学院大学がレセプトデータベースの使用に関する契約を行い、7健保組合の2005年4月から2006年3月までのレセプトデータを研究材料とした。

本データベースには、患者総数約29万人（男性：約156,000名、女性：約133,000名）、レセプト総件数は179万件を超えるデータが含まれている。

Ⅲ 分析方法

先行研究におけるシグナル抽出のアルゴリズムの基本的な枠組みは、「医薬品」と「有害事象」の1:1の組合せの総数を探索空間とし、特定の医薬品と特定の有害事象の1:1の因果関係に関するシグナルを検出している。すなわち、2×2分割表を準備して、当該医薬品・有害事象の組の報告件数が他の医薬品・有害事象の組に比べて有意に多くないかをオッズ比（ROR: Reporting Odds Ratio）を用いて判定し、多い場合にはシグナルを出すというものである。

表1 2×2分割表のセル度数

	注目する副作用	その他の副作用
注目する医薬品	n_{11}	n_{12}
その他の医薬品	n_{21}	n_{22}

表1のような2×2分割表を考えると、分子は注目する副作用群（Positive群）におけるオッズ（ n_{11}/n_{21} ）で、分母はその他の副作用群（Negative群）におけるオッズ（ n_{12}/n_{22} ）であり、その比をとったものがRORの定義である。

$$ROR = \frac{(n_{11}/n_{21})}{(n_{12}/n_{22})} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}$$

レセプトデータに含まれる多種類の情報（傷病・薬剤・処置等のアイテム）を活用し併用薬剤の相互作用、患者の持つ合併症、受けている治療や処置の内容など有害事象に至る原因事象を、より多元的に捉えたいと考えた。多種類の事象が組み合わさった原因事象を考えようとすると、先行研究で検討されている手法では組合せ爆発が起き、その探索効率は極めて悪くなる。そこで本研究では、効率的な探索を可能とするAprioriアルゴリズム^[1]を実装したEmerging Pattern^{[3][4]}に基づくシグナル探索手法をデータマイニングソフトウェアであるMUSASHI^{[6][7][8]}を利用して新たに開発した。

Emerging Pattern (EP) とは、あるクラス（群）については多頻度であるが、他のクラス（群）においては相対的に少ない頻度であるような傷病・薬剤・処置等の組み合わせ、すなわちアイテム集合としてのルールのことである。図1のように、全体のデータベースは、間質性肺炎を発症していない患者群のデータベースD1（Negative群）と間質性肺炎を発症している患者群のデータベースD2（Positive群）に分割されており、含まれるアイテム集合Xの相対頻度が異なるケースである。

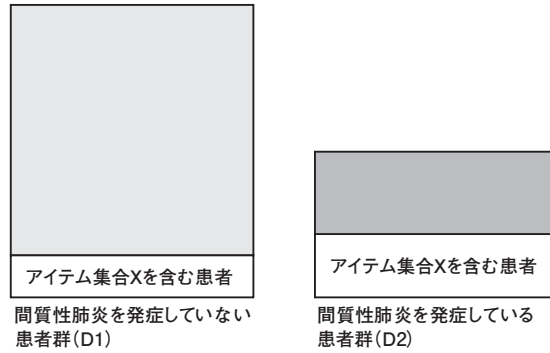


図1 Emerging Pattern (アイテム集合 X) の概念図

クラス D1、D2 におけるアイテム集合 X を考えるとき、サポート値は以下のように定義される。

D1 におけるアイテム集合 X のサポート値：

$$Supp_{D1}(X) = \frac{count(X)}{|D1|}$$

D2 におけるアイテム集合 X のサポート値：

$$Supp_{D2}(X) = \frac{count(X)}{|D2|}$$

次に D1、D2 について、アイテム集合 X の D1 に対する D2 の伸び率 (GR ; Growth Rate) は以下の通り定義される。

$$GR_{D1 \rightarrow D2}(X) = \frac{Supp_{D2}(X)}{Supp_{D1}(X)}$$

$$GR_{D1 \rightarrow D2}(X) = 0 \text{ (if } Supp_{D1}(X) = 0 \text{ and } Supp_{D2}(X) = 0)$$

$$GR_{D1 \rightarrow D2}(X) = \infty \text{ (if } Supp_{D1}(X) = 0 \text{ and } Supp_{D2}(X) \neq 0)$$

ユーザーが指定する最小サポート値 (δ) と最小 GR(ρ) について、以下の条件を示すアイテム集合 X のことを EP と定義する。

$$\begin{cases} GR_{D1 \rightarrow D2}(X) > \rho \\ Supp_{D2}(X) > \delta \quad (0 < \delta < 1) \end{cases}$$

ROR と GR の関係は、

$$ROR = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}} = \frac{n_{22} \cdot n_{+1}}{n_{21} \cdot n_{+2}} \times \frac{n_{11} \cdot n_{+2}}{n_{12} \cdot n_{+1}} = k \cdot GR \geq k\rho$$

となり、事例数が十分に大きく $n_{21} \approx n_{+1}$ および $n_{22} \approx n_{+2}$ であれば、 $k \approx 1$ で ROR と GR

はほぼ同じ値になる。このようにオッズ比 (ROR) を指標とするシグナル分析において、ルール抽出の手法として EP を適用することは論理的な妥当性があるといえる。

IV 分析結果

本研究で探索されたルールは、ルール集合の Key となる医薬品に注目し、各薬剤の 1 アイテムのルールと ROR (オッズ比)、2 アイテム以上のルールと ROR、および各薬剤の添付文書内容を比較することで論じることとした。

EP 抽出のアルゴリズムにおいては、最小サポート値を上回る全てのアイテム集合を探索することになるので、最小サポート値を低くすればするほどルール探索の負荷がかかることになる。今回の分析においては、サポート値 $\delta=0.03$ とすることによってルール検出の効率化を図った。また最小 GR(ρ) については、Positive 群 (間質性肺炎を発症した患者群) に特徴的な説明力が高いルールを見出すことを第一の目的とするため、 $\rho=10$ と高く設定した。

はじめに全ての患者群で、間質性肺炎を起こした患者とそれ以外に区分して EP 探索を行ったところ、関節リウマチ等を含む全身性の自己免疫性疾患およびその治療薬がルールとして抽出された。探索されたルールの医学的、病理解釈を考えると、自己免疫性疾患に属する患者における全身症状の一部として間質性肺炎が合併することが知られており、当たり前のルールが強い共起をもって抽出されたと考えることができた。

そこで、関節リウマチ等を含む全身性の自己免疫性疾患の病態下で、受けている治療や処置の内容など有害事象に至る原因事象を探ることとした。

その結果、薬剤の併用による有害事象の増強が推察されるルールとして、[ガスター+ロキソニン+胃潰瘍] の共起において、ROR はガスター単独の 4 から 17.6 に上昇する (表 2)。比較的一般的なこの処方も自己免疫性疾患の存在下では注意を要することが分かった。

表 2 探索されたルールと ROR 値

	Pos 群	Neg 群	
ガスター	4	23	
その他	95	2,229	ROR=4

↓

	Pos 群	Neg 群	
ガスター+ロキソニン+胃潰瘍	3	4	
その他	96	2,248	ROR=17.6

(*Pos 群 : Positive 群、Neg 群 : Negative 群)

同様に、医薬品単独の1アイテムルールと同じ医薬品を含む2アイテム以上のルールで興味深いルールをピックアップしRORを比較した(表3)。

表3 1アイテムルールと2アイテム以上のルールの比較

1アイテム	ROR		2以上のアイテムルール	ROR	
ガスター	4.0	→	ガスター/ロキソニン/胃潰瘍	17.6	併用による有害事象の増強
ガスター	4.0	→	ガスター/バクタ	23.4	併用による有害事象の増強
ウルソ	6.6	→	ウルソ/セフゾン	70.3	併用による有害事象の増強
ロキソニン	(-)	→	ロキソニン/強皮症/関節リウマチ	∞	起こりやすい病態下での使用
アタラックスP	1.4	→	アタラックスP/肺炎/関節リウマチ	35.2	根拠推定が不能
アクトネル	10.5	→	アクトネル/高脂血症	14.0	根拠推定が不能
サイトテック	16.1	→	サイトテック/関節リウマチ/プレドニン	∞	根拠推定が不能
ガスター	4.0	→	ガスター/閉塞性動脈硬化症	∞	根拠推定が不能

(-) : 抽出なし

[ガスター+バクタ]の共起において、RORはガスター単独の4から23.4に上昇した。また[ウルソ+セフゾン]の共起において、RORはウルソ単独の6.6から70.3に上昇した。これらも薬剤の併用による有害事象の増強が推察されるルールである。

次に、ロキソニンは関節リウマチなどの痛み止めとしてよく処方されるが、[ロキソニン+強皮症+関節リウマチ]のルールではROR=∞となり、強皮症など自己免疫性疾患の存在下では慎重に投与されるべき薬剤であると考えられる。

ロキソニンの添付文書には重篤な副作用として間質性肺炎の記載があるが、今回の分析でロキソニン単独ではルールとして抽出されなかった。これはロキソニンが、間質性肺炎が起こりやすい関節リウマチという病態下での使用が多いために、結果的に間質性肺炎の副作用情報が集積され添付文書に反映されている可能性もある。

また、アタラックスPは単独のルールとしてはROR=1.4であるが、[アタラックスP+肺炎+関節リウマチ]のルールではROR=35.2と極端に上昇する。アタラックスPは添付文書上、間質性肺炎の記載はない。肺炎合併下での使用で病態変化に何らかの影響を持ち、間質性肺炎発症に結びつくという新たな知見であるかもしれないが、医学的な根拠は不明である。

以上、いくつかの興味深いルールが抽出され、これらは添付文書中の副作用に関する記載の背後にある併用薬剤や合併症の存在を推測でき、実地医療での予測・予防に関するヒントを得るための情報と位置付けることができる。

V 考察とまとめ

全身性自己免疫性疾患の治療において、どのような状況で間質性肺炎が起こりやすいの

かというヒントが得られた。随伴症状の治療のために比較的よく処方されている薬剤の一部は「間質性肺炎」の発症に至る引き金を引いている可能性があることも推察できた。

発見された個々の薬剤ごとのルールは、因果関係を証明できたわけではなく、医学的根拠も不明であり、今後の科学的な検証が必要である。しかしながら、少なくとも添付文書中の副作用に関する記載の背後にある併用薬剤や合併症の存在を推測でき、実地医療での予測・予防に関するヒントを得ることができたことは価値がある。

近年では、新薬の開発に膨大な費用と時間を要し、またその成功確率も極めて低いことから、製薬企業はすでに市販されている薬剤を大切に「育薬」していくことに力を入れ始めている。

これらは本来、製薬企業の社会的役割としての薬剤の適正使用のための適切な情報提供や新しい適応症の可能性を探ることの活動であるが、質の高い医薬品使用実態調査（ファーマコビジランス）を行なっていくこと、及び市販後に実地医療での使用実態に合わせた有効性、安全性のエビデンスを求めるための大規模な臨床試験を行なっていくためには莫大なコストを製薬企業が負担しているのも事実である。

本研究では Emerging Pattern に基づくシグナル探索手法を新たに開発し、その手法をレセプトデータに適用することで重篤な副作用の発現に至る薬剤処方についての興味深い知識の発見に成功した。製薬企業がこれらの知識を活用することで、より効果的な医薬品使用実態調査（ファーマコビジランス）が可能となり、薬剤使用に関する安全性情報の質を高め、情報付加価値を加えること、及び臨床試験にかかる莫大なコスト低減することを期待でき、ビジネス上の価値が大きいと考えられる。

VI 今後の課題

本研究において、薬剤の併用順序や合併症、処置の前後関係の情報を加味した時系列の ESP（Emerging Sequential Pattern）の検討^[10]も有用であると考えたが、今回の検討からは割愛した。なぜならレセプトデータは、診療年月、薬剤処方年月といったタイムスタンプを持つが、レセプトは月毎に発生するものであり、同一（同月）レセプト内のイベントの前後関係は特定できないという性質を持つため、本研究における多種類のイベントの時系列分析は困難であると判断した。

糖尿病患者が年月を経て、腎透析に至る病態に進行するような1年以上の長期に亘る疾病のみの前後関係について分析する際には問題とならないが、今回の事例の如く薬剤投与や処置実施と有害事象発現が短期間で起こるような場合には、その薬剤が投与されたから有害事象が起こったのか、有害事象が起こったからその対処のため薬剤が投与されているのかの前後関係が明確にできるデータ構成が必要であると思われる。

次に、探索されたルールの評価手順については、データマイニング上の論点となることも多いが、基本的には探索されたルールは、ある尺度（関心度）に従って評価し、興味あるルールを選定してゆく手順を踏む。本研究では定量的な尺度は用意せず、ルール集合のKeyとなる医薬品に注目し、各薬剤を含むルールと添付文書内容を比較することで論じることとしたが、恣意的な判断が入る余地が多いといえる。

今後、より定量的で医学的知見を加味したルールの評価尺度、評価手順の開発が必要であると考えます。

〈謝辞〉

本研究を進めるにあたり、親切に、また熱くご指導頂きました関西学院大学経営戦略研究科 羽室行信助教授に深く感謝申し上げます。とくに本研究で必須のデータマイニングソフトウェアである“MUSASHI”に関して懇切丁寧に教えて頂いたこと、また論文作成において主査として多大なるご助言を頂きましたことに心より感謝致します。

本研究の分析対象であるレセプトデータのご提供、およびデータ関連の様々なご助言を頂きました株式会社日本医療データセンター（JMDC）代表取締役社長 木村真也氏、ヘルスケア事業統括部 福原美穂氏、石原寛子氏、システム本部 林哲氏に深く感謝致します。

また、ともに学び、切磋琢磨した羽室ゼミの学友の皆様、2年間文句も言わず支えてくれた家族に感謝致します。

最後に、社会人としてこの学舎で存分に学べたことを誇りに思い、本大学院研究科の学生や活動を陰で支えて頂いた山本昭二教授に敬意を表します。

参考文献

- [1] R. Agawal and R. Srikant “Fast algorithms for mining association rules.” *Proc. of the 20th VLDB Conf.* 207–216, 1994.
- [2] Michael J.A. Berry, Gordon S. Linoff 『データマイニング手法』海文堂出版, 2005.
- [3] G. Dong and J. Li “Efficient mining of emerging patterns: Discovering trends and differences.” *In Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99)* 43–52, San Diego, CA, Aug. 1999.
- [4] G. Dong et al. “CAEP: Classification by Aggregating Emerging Patterns” *Int'l Confidence on Discovery Science*, 30–42, 2005.
- [5] 藤田利治他「医薬品の副作用自発報告によるシグナル検出の実用化に向けての検討」厚生労働科学研究補助金分担研究報告書, 2004.
- [6] 羽室行信他「MUSASHI でらくらくデータマイニング」*Software Design*, 2003.
- [7] 羽室行信他「大規模ビジネスデータからの知識発見システム:MUSASHI」*人工知能学会誌*, 20 (1), 2005.
- [8] 羽室行信他「知識発見支援ソフトウェア:MUSASHI」*オペレーションズ・リサーチ*, 50 (9), 654–659, 2005.
- [9] 平山佳伸「医薬品情報とファーマコビジランス」*Pharmavision*, 9 (7), 2005.
- [10] 石倉寛子「レセプトデータへのマイニング技術の適用による疾病予測支援システムの構築」大阪産

業大学大学院工学研究科 アントレプレナー専攻 修士論文, 2006.

[11]厚生労働省「厚生労働白書」平成 18 年度版

[12]Michael Lawson ed. 「製薬産業の未来」MIT technology insider 9 月号, 2003.

(www.ntt.co.jp/sclab//images/ilp/mit_pdf/200309.pdf) [2007. 01. 13 アクセス]

[13]元田浩, 鷺尾隆「データマイニング展望」システム/制御/情報, 46 (4), 169-176, 2002.

[14]日本医療データセンター資料「レセプト (診療報酬明細書) データの活用事例」Spotfire Customers Conference, 11-12 Nov. 2004.

[15]岡本悦司, 木村真也「レセプトデータを使用した医薬品有害事象検出データマイニング手法の開発」薬剤疫学, 10 (1), 15-23, 2005.