厚生労働科学研究費補助金（医薬品・医療機器等レギュラトリーサイエンス総合研究事業）
コンパニオン体外診断用医薬品の臨床性能試験の在り方に関する再帰的研究
平成23年度分担研究報告書

レギュラトリーサイエンスに基づくコンパニオン体外診断用医薬品の臨床性能試験の在り方に関する研究
—審査過程の明確化を目指した審査報告書のテキストマイニング—

分担研究者　横井　英人（香川大学医学部附属病院　医療情報部　教授）
研究協力者　小野　大樹（香川大学医学部　大学院博士課程）

研究要旨

近年のテーラーメード医療（個別化医療）の発展により、ヒト疾患の診断を中心とするバイオマーカーの累計特許出願件数は増加傾向にある。1995年1月1日から2005年12月31日までに累積で3,890件の出願があり、日本は、欧州と米国同様に増加傾向である[1)]。このような中、今後はバイオマーカーをあらかじめ調べるための体外診断薬（以下、コンパニオン診断薬）とその診断薬が不可欠な分子標的医薬品の開発・申請が急増する可能性が高い。しかしながら、このコンパニオン診断薬の位置づけや承認審査の基準は未だ明確ではない。

本研究では、コンパニオン診断薬の承認審査の基準を明確にするための知識ベースの構築に向けて、まずはそのパイロットスタディとして、コンパニオン診断薬に関連の深い抗癌剤、中でも分子標的薬の承認審査報告書を対象にテキストマイニングを行った。

A. 研究目的

本研究では、コンパニオン診断薬の承認審査に関する知識ベースの構築を目指している。今回は、コンパニオン診断薬を検討する前に、その対象となる抗癌剤を検討する必要があると考え、パイロットスタディとして抗癌剤の中でも分子標的薬の承認審査報告書を対象にテキストマイニングを行い、テキストマイニングツールの評価・検討を行った。

B. 研究方法

1．対象

対象には、効能効果に偏りが少ない分子標的薬10種類を選定し、インターネット上に公開されている新薬承認審査報告書データベース[2)]から、各々の審査報告書を抽出した。表1に抽出した分子標的薬の特徴を示す。

2．方法

2.1 形態素解析ツール

形態素解析とは、文書を構成する文字列を単語に分割し、各単語に品詞や語形変化などの情報を与える処理のことである。

近年、数多くの形態素解析ツールが出てきているが今回の研究では、審査報告書の形態素解析だけでなく、その内容の論理展開までを解析対象としているため、形態素解析ツールには、インターネット広告やマーケティングの領域において感性分析[4)]の手法で定評のある、日本語解析エンジン「なずき」[3)]（以下、「なずき」）を採用した。

2.2 感性分析

感性分析[4,5]は、主にマーケティングの分野において、ある商品に対する印象や評価が書かれているアンケートやブログ・ツイッターの分析などに使われている。

従来の手法では、意図を理解するには結局人間が文章を読む必要があったが、今回採用した「なずき」の感性分析によると、単語を照合するのではなく、人間の感性を81種類に分類し、それぞれの感性ごとに実際にはどのような文章表現が用いられるのかというパターン辞書を使って意見を見つけ出すことが可能5)となる。具体例を示すと、「医薬品の効果が確認できたので良かった」は「医薬品に対する満足」、「医薬品の効果が確認できたら良かったのに」は「医薬品に対する要望」として検出することができる。

2.3 ベクトル空間モデル

ここでは、ベクトル空間モデル 6,7)について述べる。今回対象とする審査報告書の集合を **D** とし, 各審査報告書を **d1, d2, …, dj, …, d10** とおく。また、集合 **D** から「なずき」によって抽出した32,390個の単語を **w1, w2, …, wi, …, w32390** とする。次に、ある審査報告書 **dj** に現れる単語 **wi** に対する重みを **αij** とおく。このとき **dj** を次のようなベクトル

$$d_j = [\alpha_{1j}\ \alpha_{2j}\ \dots\ \alpha_{ij}\ \dots\ \alpha_{32390j}]$$

で表現し、これをある分子標的薬 **j** の審査報告書のベクトルとする。

また対象の審査報告書の集合全体は，次のような32390×10行列 **D**

$$\mathrm{D} = [d_1 d_2 \dots d_{10}] = \begin{bmatrix} \alpha_{1\,1} & \cdots & \alpha_{1\,10} \\ \vdots & \ddots & \vdots \\ \alpha_{32390\,1} & \cdots & \alpha_{32390\,10} \end{bmatrix}$$

で表現する事ができる。なお、索引語の重み付けの計算には、先行研究[8,9]でも採用されている TF*IDF 法を用いることとし、重み **α** は

$$\alpha_{ij} = l_{ij} g_i$$

で表し、**lij** と **gi** は以下のように定義する。

$$l_{ij} = \log(1 + f_{ij})$$
$$g_i = \log\left(\frac{n}{n_i}\right)$$

ここで、**fij** とは索引語 **wi** の審査報告書 **dj** における出現頻度である。また、**n** は対象とする審査報告書の種類であり、**ni** は索引語 **wi** が出現した審査報告書の数を示している。

また、審査報告書のベクトル間の類似度は、以下の式を用いてコサイン類似度で表現する。

$$\cos\theta = \frac{dj \cdot dk}{\|dj\|\|dk\|} = \frac{\sum_{i=1}^{32390} dj_i dk_i}{\sqrt{\sum_{i=1}^{32390} dj_i^2} \cdot \sqrt{\sum_{i=1}^{32390} dk_i^2}}$$

C. 結果

1．感性分析結果

まず、セツキシマブ（商品名：アービタックス）の審査報告書に対して行った「なずき」による感性分析の判定結果とその結果に対する専門家による評価を表 **2** に示す。

ここで、専門家による評価の「肯定」は、セツキシマブに対する肯定的な評価を示し、「否定」はセツキシマブに対する否定的な評価を示している。次に、「肯定・否定」はセツキシマブに対して肯定的とも否定的とも言えない評価、「不適」はセツキシマブの評価に直接関係のない記述、「不明」はセツキシマブの評価に直接関係あるか不明な記述であったものをそれぞれ集計して示している。

結果を見ると、「なずき」による判定と専門家の評価が一致していたものは、肯定で39件（10%）、否定で55件（11%）となった。一方で、「肯定・否定」や「不適」と評価されたものが半数以上あり、その他にも「なずき」の判定では「肯定的」であっても、専門家の評価が「否定的」であったものや、その逆の事例も散見された。

今後、本研究の大目的である、承認に至るまでの審査の論理展開の定型化・類型化を試みるためには、「なずき」の感性分類のパターン辞書やルールを審査報告書に特化した形に追加・修正をしていく必要性があると思われた。

2．単語の重み付けランキングと類似度

次に、各分子標的薬と関連が深いと思われる単語と分子標的薬間の類似度を示す。

なずきによる形態素解析後に、前述のTF*IDF法を用いて重み付けし、各分子標的薬の審査報告書のベクトルを算出し、それぞれの形態素を重み順に並べ替えることで、関連が強いと思われる単語を抽出した。表3に各分子標的薬毎の上位15位までの単語を示す。

さらに、各分子標的薬同士の類似度を表4に示す。その結果を見ると、10種の分子標的薬間の中で一番類似度の数値が大きかったのは、「スーテントとネクサバール」の(0.14)であり、順に「アービタックスとアバスチン」の(0.118)、「ハーセプチンとリツキサン」の(0.101)と続いていた。「スーテントとネクサバール」は共に、効能効果に「根治切除不能又は転移性の腎細胞癌」の適応を有しており、標的も「VEGFR-TKI」と同一であり、薬剤のプロファイルからも類似性が高い。次いで「アービタックスとアバスチン」においては、アービタックスは「EGFR 陽性の治癒切除不能な進行・再発の結腸・直腸癌」、アバスチンは「治癒切除不能な進行・再発の結腸・直腸癌」の適応を有しており、癌の部位が結腸・直腸であることから、同様に関連性が高い薬剤であった。

D. 考察

1．マイニング結果について

1.1 感性分析結果

今回、用いたマイニングツール「なずき」の感性分析は、主として商業活動に於けるフリーテキストからのナレッジベース構築に用いることを目的としている。具体的にはコールセンターへの顧客からの訴えをテキストに起こし、その内容を分析し、対象商品（サービス）などについて「どのような点」が「良かった」のか（肯定）、「どのような点」が「悪かった」のか（否定）という観点で集計可能である。

フリーテキストから正規的な情報（明確な形の「属性」とその「属性値」）を抽出できれば、大量のテキストデータが発生する大手のコールセンターでは有効に作用するツールとなりうるであろう。しかし、今回、実際に同ツールの分析結果の一例を、原文をたどりながら検証した結果、「Aを評価する」を「A に良い評価を与える」という意味に取っていた。実際に審査報告書での「Aを評価した」という記述は、「Aの評価を行った」という意味であったので、これは肯定でも否定でもない文として、専門家の評価としては「不適」とされた。このような

例が散見された他、疾患名や症候名に対して、同ツールが「悪い」評価と判定する根拠としているきらいがあった。多くの有害事象は、このような疾患名や症候名で示されるので、その記述自体が否定的な記述と判定されたが、仮に有害事象が発生していても、それが同系薬・同効薬に較べて少なければ、同薬剤の安全プロファイルは肯定的な評価を受けるべきである。このような例を鑑みるに、審査報告書に於ける薬剤の評価ロジックを、現時点の同ツールでの肯定・否定の検出アルゴリズムで十分に検出することは困難であろうと考えられた。

2．単語の重み付けランキングと類似度

また TF*IDF 法を用いた単語の重み付けを使用して、各分子標的薬の類似度を検討した結果は、概ね効能効果の類似性と相関があるように思われた。このことは、文書内容のサマライズに有効であるという結果を多く持つ同手法が、一定の効果を示したと考えることが出来る。しかし審査報告書では、当該薬品名称がそもそもがコード名で呼称されていることが多い他、その薬品の評価を考えるに大変重要な臨床評価も、企業が振り付けた試験名称（これもほとんどがコード名）で記載されている。このことにより、「○○試験は～のような試験である」というようなメタ情報がなければ、実際のマイニング結果から、これ以上薬に関する評価情報を抽出することは出来ない。この点も次年度以降、検討が必要である。

E. 結論

本年度は、コンパニオン診断薬の承認審査に関する知識ベースの構築に向けて、まずは、抗癌剤特に分子標的薬の審査報告書を対象にテキストマイニングを試みた。審査報告書の形態素解析やベクトル空間モデル・TF*IDF 法による重み付けによって、各分子標的薬の特徴となりうる単語を抽出できた。今後の課題は、マイニングツールを審査報告書向けにチューニングすることである。具体的には、

- 同義語の正規語への統一化の処理
- 従来の辞書に存在しない単語の登録
- 感性分析に用いる判定パターンやルールの追加・修正

を中心に、進めていく必要があると思われた。

F. 健康危険情報

特になし。

G. 研究発表

1．学会発表

1) 小野大樹, 尾崎哲夫, 池田正行, 横井英人, コンパニオン診断薬の承認審査に関するナレッジベースの構築に向けたテキストマイニング技術活用の検討, 第31回医療情報学連合大会/第12回日本医療情報学会学術大会, 2011

H. 知的財産権の出願・登録状況

1．特許取得

なし。

2．実用新案登録

なし。

3．その他

なし。

参考文献

[1] 鳥山裕司.医薬関連バイオマーカーの特許出願動向にみる日本の課題.政策研ニュース 2008；26：22-26.

[2] JAPIC 日本の新薬審査報告書DB.http://www.shinsahoukokusho.jp/dar_us/dar/search/usDarSearch.jsp.

[3] 青江順一,結束雅雪.継続こそ力:「なずき」開発物語(特論-4,<特集>イノベーションが生まれたルーツを探る).品質 2007；37(3)：246-251.

[4] Masao Fuketa, Yuki Kadoya, El-Sayed Atlam, et al. AMethod of Extracting and Evaluating Good and Bad Reputations for Natural Language Expressions. Information Technology & Decision Making 2005； Vol.4, No.2：177-196.

[5] 株式会社NTTデータ.なずきエモーションアナライザ Ver.1.4 使用説明書.

[6] Salton G, Wong A, Yang C S.A Vector Space Model for Automatic Indexing. CACM 1975 ；18：613-620.

[7] 北研二、津田和彦、獅々堀正幹.ベクトル空間モデルに基づく文書検索.情報検索アルゴリズム.共立出版,2002：50-64.

[8] 鈴木隆弘, 小野大樹, 横井英人, 井宮淳, 高林克日己.退院サマリーのテキストマイニングにおけるエントロピー法とｔｆ×ｉｄｆ法の比較.医療情報学 2005；25(3)：173-180.

[9] 小野大樹, 高林克日己, 鈴木隆弘, 横井英人, 井宮淳, 里村洋一.テキストマイニングによる退院サマリー自動分類の試み.医療情報学 2004；24：35-44.

表 1 今回対象とした 10 種類の分子標的薬

一般名	商品名	剤形	効能効果	標的	主な副作用(上位3つ)	重大な副作用(上位3つ)
セツキシマブ(遺伝子組換え)	アービタックス	注射液	EGFR陽性の治癒切除不能な進行・再発の結腸・直腸癌	EGFR	ざ瘡(87.2%)、発疹(61.5%)、食欲不振(56.4%)	重度のinfusion reaction、重度の皮膚症状、間質性肺炎患
リツキシマブ(遺伝子組換え)	リツキサン	注射液	1.CD20陽性のB細胞性非ホジキンリンパ腫 2.インジウム(111In)イブリツモマブ チウキセタン(遺伝子組換え)注射液及びイットリウム(90Y)イブリツモマブチウキセタン(遺伝子組換え)注射液投与の前投与	CD20	発熱(64.3%)、悪寒(34.4%)、そう痒(21.7%)	アナフィラキシー様症状、肺障害、心障害
メシル酸イマチニブ	グリベック	経口	1.慢性骨髄性白血病 2.KIT(CD117)陽性消化管間質腫瘍 3.Ph'フィラデルフィア染色体陽性急性リンパ性白血病	Bcr-Abl-TKI	嘔気(45.7%)、好中球減少症(42.9%)、血小板減少症(40.0%)	骨髄抑制、出血、消化管穿孔
ボルテゾミブ	ベリケイド	注射液	再発又は難治性の多発性骨髄腫	プロテアソーム	貧血(73.5%)、リンパ球数減少(64.7%)、白血球数減少	肺障害、心障害、末梢性ニューロパシー
ベバシズマブ(遺伝子組換え)	アバスチン	注射液	1.治癒切除不能な進行・再発の結腸・直腸癌 2.扁平上皮癌を除く切除不能な進行・再発の非小細胞肺癌	VEGF	好中球減少症(18.8%)、白血球数減少(18.5%)、高血圧(14.6%)	ショック、アナフィラキシー様症状、消化管穿孔、瘻孔
トラスツズマブ(遺伝子組換え)	ハーセプチン	注射液	1.HER2過剰発現が確認された転移性乳癌 2.HER2過剰発現が確認された乳癌における術後補助化学療法 3.HER2過剰発現が確認された治癒切除不能な進行・再発の胃癌	HER2	乳癌:悪寒(4.5%)、肺痛(3.6%)、発熱(3.5%) 胃癌:悪心(63.3%)、好中球減少症(53.4%)、嘔吐(43.9%)	心障害、アナフィラキシー様症状、間質性肺炎・肺障害
ソラフェニブトシル酸塩	ネクサバール	経口	1.根治切除不能又は転移性の腎細胞癌 2.切除不能な肝細胞癌	VEGFR-TKI	リパーゼ上昇(58.6%)、手足症候群(55.2%)、アミラーゼ上昇(40.7%)	手足症候群、剥脱性皮膚炎、皮膚粘膜眼症候群(Stevens-Johnson症候群)
スニチニブリンゴ酸塩	スーテント	経口	1.イマチニブ抵抗性の消化管間質腫瘍 2.根治切除不能又は転移性の腎細胞癌	VEGFR-TKI	血小板減少(91.4%)、白血球減少(85.2%)、皮膚変色(82.7%)	骨髄抑制、感染症、高血圧
ゲムツズマブオゾガマイシン(遺伝子組換え)	マイロターグ	注射液	再発又は難治性のCD33陽性の急性骨髄性白血病	CD33	発熱(95.0%)、血小板減少(95.0%)白血球減少(92.5%)	infusion reaction、重篤な過敏症、血液障害(骨髄抑制等)
ゲフィチニブ	イレッサ	経口	手術不能又は再発非小細胞肺癌	EGFR-TKI	発疹(62.7%)、下痢(49.0%)、そう痒症(49.0%)	急性肺障害、間質性肺炎、重度の下痢、脱水

表 2 「なずき」の感性分析による判定と専門家による評価

なずきによる判定		専門家による評価					
判定結果	判定数	肯定	否定	肯定・否定	不適	不明	合計
肯定	410	39(10%)	15(4%)	12(3%)	344(84%)	0	410
否定	508	55(11%)	65(13%)	225(44%)	160(31%)	3(1%)	508

表 3 各分子標的薬の単語とその重み順

順位	アービタックス		アバスチン		イレッサ		グリベック		スーテント	
	INDEX	重み	INDEX	重み	INDEX	重み	INDEX	重み	INDEX	重み
1	CPT－11	11.080	XELOX	9.972	CSS	7.220	ABL	5.906	SU012662	12.057
2	EGFR陽性	9.847	AVF2107G	9.783	MS法	6.898	GIST	5.675	不忍容	8.376
3	IMCL	9.098	AVF2192G	8.606	プラチナ系	6.655	BCR	5.059	総薬物	7.980
4	CPT－11群	9.053	抗ベバシズマブ抗体	8.494	褐色製剤	6.655	CML患者	5.059	スニチニブ	7.589
5	CPT－11併用群	8.660	FOLFOX4	8.376	黄色製剤	6.524	CML	4.900	GIST	7.428
6	維持投与量250MG	8.186	LV群	7.753	褐色錠	6.524	GIST患者	4.900	報告書SU011248	7.412
7	CTG	8.120	血清中ベバシズマブ濃度	7.753	500MG群	6.420	KITチロシンキナーゼ活性	4.788	PDGFR	7.131
8	初回投与量400MG	8.120	AVF0780G	7.673	AGP濃度	6.236	以下同様	4.788	リンゴ酸塩	7.117
9	抗セツキシマブ抗体	8.051	LOT	7.673	700MG	5.906	600MG群	4.528	腎細胞癌	7.053
10	CA	7.784	国際電気標準会議	7.589	HPLC	5.906	欧米人GIST患者	4.481	IFN－A群	6.780
11	CPT－11単独群	7.753	LV	7.580	最小最小二乗	5.906	既承認申請資料	4.481	スニチニブリンゴ酸塩	6.780
12	試験番号IMCL	7.589	国内JO18157	7.412	225MG	5.722	効果持続期間	4.481	RET	6.524
13	ILD	7.412	第二群	7.010	国際共同	5.722	国内外臨床試験	4.481	カプセル剤	6.394
14	事象名	7.412	DG	6.955	無酸症	5.521	400MG群	4.441	GIST患者	6.359
15	CS	7.318	IFL	6.479	雄イヌ	5.521	中間集計	4.358	1日1回50MG	6.236
順位	ネクサバール		ハーセプチン		ベルケイド		マイロターグ		リツキサン	
	INDEX	重み	INDEX	重み	INDEX	重み	INDEX	重み	INDEX	重み
1	AUC0－12	7.753	転移性乳癌	8.914	JPN101	9.466	HP67．6	11.931	CD20	9.008
2	手足皮膚反応	7.589	維持量	8.376	多発性骨髄腫	9.269	CD33	11.785	非ホジキンリンパ腫	8.865
3	50MG錠	7.412	FISH法	7.589	プロテアソーム活性	8.120	カリケアマイシン誘導体	11.172	マントル細胞リンパ腫	7.980
4	相併用試験	6.780	負荷量	7.318	デキサメタゾン群	8.051	バート	10.488	低悪性度	7.907
5	1回400MG	6.655	HERCEP	7.117	主公表	7.589	VOD	10.175	濾胞性	7.589
6	200MG錠	6.655	AC療法	6.898	MM	6.780	殺細胞活性	9.692	ろ胞性B細胞性	7.220
7	MOTZER	6.394	PTX単独群	6.236	RPMI8226	6.780	CMA－676	8.961	ラージ	6.780
8	プロセス	6.298	IHC法	6.197	血液中20	6.655	9MG	8.865	SMALL	6.655
9	リスク分類	6.236	初回負荷量	5.906	プロテアソーム	6.524	寛解後療法	8.765	375MG	6.524
10	腎細胞癌	6.162	HT	5.722	プロテアソーム活性阻害率	6.524	カリケアマイシン	8.660	試験番号IDEC	6.394
11	実施期間20	6.090	過剰発現	5.670	個人輸入症例	6.394	総カリケアマイシン誘導体	8.436	B細胞	6.264
12	PUGH分類	6.077	NEU遺伝子	5.521	ボルテゾミブ	6.077	CRP例	8.376	B細胞性	6.236
13	本薬400MG	5.977	RIII	5.521	起立性低血圧	6.077	寛解導入療法	8.120	1回375MG	5.906
14	AUCNORM	5.722	オープン試験	5.521	CHT	5.906	抗CD33	7.907	CHOP療法	5.521
15	コホートあたり	5.722	重度1件	5.521	プロテアソーム阻害活性	5.906	非結合カリケアマイシン誘導体	7.907	リンパ腫	5.521

表 4 分子標的薬間の類似度

類似度(COSθ)	アービタックス	アバスチン	イレッサ	グリベック	スーテント	ネクサバール	ハーセプチン	ベルケイド	マイロターグ	リツキサン
アービタックス	-	0.118	0.061	0.039	0.096	0.084	0.060	0.070	0.061	0.045
アバスチン	0.118	-	0.054	0.034	0.086	0.089	0.072	0.074	0.073	0.040
イレッサ	0.061	0.054	-	0.058	0.076	0.085	0.064	0.058	0.055	0.043
グリベック	0.039	0.034	0.058	-	0.062	0.058	0.035	0.032	0.032	0.026
スーテント	0.096	0.086	0.076	0.062	-	0.140	0.047	0.080	0.059	0.027
ネクサバール	0.084	0.089	0.085	0.058	0.140	-	0.046	0.077	0.054	0.026
ハーセプチン	0.060	0.072	0.064	0.035	0.047	0.046	-	0.056	0.064	0.101
ベルケイド	0.070	0.074	0.058	0.032	0.080	0.077	0.056	-	0.074	0.040
マイロターグ	0.061	0.073	0.055	0.032	0.059	0.054	0.064	0.074	-	0.059
リツキサン	0.045	0.040	0.043	0.026	0.027	0.026	0.101	0.040	0.059	-