別添４

年度	申請数	結果	NTAP申請技術		薬事承認 or 上市年月	機器区分	NTAP承認償還額	償還年度
FY2003	5	承認 1 保留 1 却下 2 取り下げ 1	Drotrecogin Alfa (Activated) – Xigris	医薬品（敗血症）	2001年11月	–	$3,400	FY2003, FY2004
			BMPs for Spinal Fusions	医薬品（骨形成タンパク）	2002年07月	–	(pending)	（翌年度再申請）
			Zyvox	医薬品（グラム陽性菌抗生剤）	2000年04月	–	×	–
			Renew Radio Frequency Spinal Cord Stimulation Therapy	高周波脊髄刺激治療	1999年07月	?	×	–
FY2004	2	承認 1 却下 1	BMPs for Spinal Fusions [2]	医薬品（脊髄固定術用骨形成タンパク）	2002年07月	–	$8,900	FY2004, FY2005
			GLIADEL Wafer	医薬品（カルムスチン徐放剤）	1996年09月	–	×	–
FY2005	10	承認 2 却下 8	BMPs for Tibia Fractures	医薬品（脛骨骨折用骨形成タンパク）	2004年04月	–	×	–
			Norian Skeletal Repair System (SRS) Bone Void Filler	骨空隙充填剤	2001年12月	510(k)	×	–
			InSync Defibrillator System	除細動心臓再同期システム	2002年06月	PMA	$16,262.50	FY2005 only
			GliaSite Radiation Therapy System	腔内近接照射システム	2001年04月	510(k)	×	–
			Natrecor – Human B-Type Natriuretic Peptide	医薬品（B型ナトリウム利尿ペプチド）	2001年08月	–	×	–
			Kinetra Implantable Neurostimulator for Deep Brain Stimulation	埋め込み型神経刺激装置	2003年12月	PMA	$8,285	FY2005, FY2006
			Intramedullary Skeletal Kinetic Distractor (ISKD)	内部式骨延長術	2001年05月	510(k)	×	–
			Acticon Neosphincter	便失禁治療インプラント	2001年12月	PMA	×	–
			TandemHeart Percutaneous Left Ventricular Assist System	左心補助システム	2000年11月	?	×	–
			Aquadex System 100 Fluid Removal System (System 100)	過剰体液除去システム	2002年06月	510(k)	×	–
FY2006	8	承認 2 却下 6	BMPs for Tibia Fractures [2]	医薬品（脛骨骨折用骨形成タンパク）	2004年04月	–	×	–
			Aquadex System 100 Fluid Removal System (System 100) [2]	過剰体液除去システム	2002年06月	510(k)	×	–
			CHARITE Artificial Disc	人工椎間板	2004年10月	PMA	×	–
			Endovalsular Graft Repair of the Thoracic Aorta	胸部エンドバスキュラーグラフト	2005年03月	PMA	$10,599	FY2006, FY2007
			Restore RechargeableImplantableNeurostimulator	再充電式埋め込み型神経刺激装置	2004年04月	PMA	$9,320	FY2006, FY2007
			Safe-Cross® Radio Frequency Total Occlusion Crossing System	高周波完全閉塞穿通システム	2004年01月	510(k)	×	–
			Trident Ceramic Acetabular System	人工関節	2003年04月	PMA	×	–
			Wingspan Stent System With Gateway PTA Balloon Catheter	PTAバルーン＝ステントシステム	–	–	×	–
FY2007	3	承認 1 却下 1 取り下げ 1	C-Port Distal Anastomosis System	血管吻合システム	2005年11月	510(k)	×	–
			NovoSeven for Intracerebral Hemorrhage	医薬品（遺伝子組み換え第VIIa因子）	(–)	(–)	(取り下げ)	(–)
			X STOP Interspinous Process Decompression System	腰椎伸展ストッパー	2005年11月	PMA	$4,400	FY2007 only
FY2008	1	却下 1	Wingspan Stent System With Gateway PTA Balloon Catheter [2]	PTAバルーン＝ステントシステム	2005年08月	HDE	×	–
FY2009	4	承認 1 却下 3	CardioWest Temporary Total Artificial Heart System (CardioWest TAH-t)	暫定使用全置換型人工心臓	2004年10月	PMA	$53,000	FY2009, FY2010, FY2011
			Emphasys MedicalZephyr Endobronchial Valve (Zephyr EBV)	気管支内バルブ	–	–	×	–
			Oxiplex	脊髄神経根に適用される吸収性ゲル	–	–	×	–
			TherOx Downstream System	高度酸素化血液供給システム	–	–	×	–
FY2010	6	承認 1 却下 1 取り下げ 4	The AutoLITT System	MRI下レーザー焼灼システム	(–)	(–)	(取り下げ)	(–)
			CLOLAR (Clofarabine) Injection	医薬品（小児急性リンパ芽球性白血病治療薬）	(–)	(–)	(取り下げ)	(–)
			LipiScan Coronary Imaging System	近赤外冠動脈イメージングシステム	2008年04月	510(k)	×	–
			Spiration IBV Valve System	気管支鏡下肺気腫・気胸治療システム	2008年10月	HDE	$3,437.50	FY2010, FY2011
			TherOx Downstream System [2]	高度酸素化血液供給システム	(–)	(–)	(取り下げ)	(–)
			Emphasys MedicalZephyr Endobronchial Valve (Zephyr EBV) [2]	気管支内バルブ	(–)	(–)	(取り下げ)	(–)
FY2011	5	承認 1 却下 2 取り下げ 2	TachoSil	組織接着用シート	(–)	(–)	(取り下げ)	(–)
			Dynesys Dynamic Stabilization System	脊椎固定システム	(–)	(–)	(取り下げ)	(–)
			Auto Laser Interstitial Thermal Therapy (AutoLITT) System [2]	MRI下レーザー焼灼システム	2009年05月	510(k)	$5,300	FY2011, FY2012, FY2013

年度	申請数	結果	NTAP申請技術		薬事承認 or 上市年月	機器区分	NTAP承認償還額	償還年度
			LipiScan Coronary Imaging System [2]	近赤外冠動脈イメージングシステム	2008年04月	510(k)	×	–
			LipiScan Coronary Imaging System With Intravascular Ultrasound (IVUS)	血管内超音波下近赤外冠動脈イメージングシステム	2008年04月	510(k)	×	–
FY2012	3	却下 2 取り下げ 1	Champion HF Monitoring System	心不全モニタリングシステム	(–)	(–)	(取り下げ)	(–)
			AxiaLIF 2L+ System	埋込脊椎固定システム	2010年01月	510(k)	×	–
			PerfectCLEAN With Micrillon	塩素を含む清浄用繊維	–	–	×	–
FY2013	6	承認 3 却下 1 取り下げ 2	Glucarpidase (Trade Brand Voraxaze)	医薬品(メトトレキサート腎毒性低下薬)	2012年01月	–	$45,000	FY2013
			DIFICID (Fidaxomicin) Tablets	医薬品(クロストリジウム・ディフィシル関連下痢症治療薬)	2011年05月	–	$868	FY2013
			Zilver PTX Drug Eluting Stent	下肢用DES	–	–	×	–
			Zenith Fenestrated Abdominal Aortic Aneurysm (AAA)	カスタムメイドAAA血管内グラフト	2012年04月	PMA	$8,171.50	FY2013

別添５

テキストマイニングにより
医療機器添付文書から抽出される
リスク認知の国際調査

平成 25(2013)年 3 月

ＮＫＳＪリスクマネジメント株式会社

リスクエンジニアリング事業本部　藤原俊明

目次

＜記号解説＞

[1,2]: 脚注

[1],[2]: 参考文献

(1),(2): 引用データ及びプログラム

1　緒言

1.1 研究の背景･目的

社会を取り巻くリスクに関する正確な情報を、行政、専門家、企業、市民などのステークホルダーである関係主体間で共有し、相互に意思疎通を図ることをリスクコミュニケーションと呼ぶ[1]。リスクコミュニケーションは、特に、大規模な自然災害が発生した後に、しばしばその在り方について議論される。例えば、東日本大震災後の政府のリスクコミュニケーションに対し、様々な議論がなされたことは記憶に新しい。現在、厚生労働省[2]や消費者庁[3]といった行政が、市民との適切なリスクコミュニケーションの在り方を検討すべく、意見交換会等を開催していることからも、行政がリスクコミュニケーションを重視していることがうかがえる。

適切なリスクコミュニケーションを考える上で、リスクを市民に伝達するという性質上、市民がリスクをどのように受け止めるか、すなわち、市民のリスク認知を踏まえることが重要といえよう[1][2]。

リスク認知の代表的研究として、Slovic が作成したアメリカ人のリスク認知地図が挙げられる[3]。Slovic は、未知性(unknown)と恐ろしさ(dread)の 2 因子を評価軸として、原子力事故、遺伝子工学、核兵器といたリスクに対してアメリカ人の 2 因子の感情の誘起度合いをアンケートによって明らかにし、この結果を、相対位置プロットにより 2 次元平面上で可視化した。この Slovic のような研究アプローチは、日本人のリスク認知の比較研究にも応用されている[4]。

Slovic のような研究アプローチは、あるリスク―例えば、原子力事故―に対する感じ方を日本人と米国人で、2 次元平面上の相対位置を比較することで把握するのに優れた手法である。例えば、日米大学生の比較において、遺伝子研究というリスクは、米国人より日本人のほうがより恐ろしさを感じることを容易に把握

---

[1] リスク・コミュニケーション. In Wikipedia: *The Free Encyclopedia.*
[2] 厚生労働省: "*食品の安全に関するリスクコミュニケーション*", http://www.mhlw.go.jp/topics/bukyoku/iyaku/syoku-anzen/riskcom/
[3] 消費者庁: "*食品と放射性物質に関するリスクコミュニケーション等について*", http://www.caa.go.jp/jisin/r_index.html#ris-top

することができる[4]。しかしながら、これらは、あくまで未知性と恐ろしさという2軸のみの評価である。リスク認知をさらに深く突き詰めるたい場合には、より多くの因子を踏まえることを考える必要がある。しかしながら、従来のアプローチの多くはアンケートによるものである。この場合、評価軸を増やすことにより質問数が線形的（リスク項目 x 評価軸）に増加するため、多くの因子を踏まえたアンケートを実施することは現実的には容易ではない。

本研究では、この問題を避けるため、アンケートによるアプローチではなく、任意に作成された文章群に対するテキストマイニングによってリスク認知に関する知見を得ようとするものである。テキストマイニングとは、主にコンピュータを利用することにより、任意の文章の出現する単語や文節で区切り、それらの出現頻度や、共出現の相関、出現傾向、時系列等を解析することで有用な情報を取り出す、テキストデータの分析手法を指す[5]。従来、テキストマイニングによりテキストデータを解析することは容易ではなかった[6]。しかし近年では、テキストマイニングに必要かつ有用なツールがインターネット上で多く公開されている。これらツールを組み合わせることで、目的に応じたテキストマイニングが可能となる。本報告書は、テキストマイニングによりリスクを表す語（病名等）を抽出・解析することで、リスク認知についての基礎的知見を得ることを目的とし、独自のプログラム環境を構築して実際に解析した結果を報告するものである。

ここで、テキストデータして医療機器の添付文書を選択した。また、日本人のリスク認知を、米国人のリスク認知との比較によって浮き彫りにすべく、米国の医療機器の Labeling を比較対象のテキストデータとした。当初、医療機器の添付文書/Labeling をテキストマイニングすることで、日本人と米国人のリスク認知に関する差異を明らかできるかどうかは当然明らかではなかったため、事前に簡単な比較調査を実施している。この結果を以下に紹介する。

比較したのは、任意に選んだ事業者が日本及び米国で販売したステントグラフトの添付文書と Labeling である。リスクに係る固有名詞として、添付文書の「不具合・有害事象」に記載されている事象を、また、Labeling の「Adverse Event」

---

[4] http://hse-risk-c3.or.jp/rc/rc2145.html
[5] テキストマイニング. In Wikipedia: *The Free Encyclopedia.*
[6] たとえば「わかち書き」と呼ばれる単語と単語の境目（境界判別）をすることでさえ、コンピュータに判断させるのは困難であった。

を比較した。同じ製品であれば同じ事象が記載されていると想定された。しかしながら、実際には日本の添付文書には、Labeling に含まれない 14 の有害事象があった。一方、Labeling には、添付文書にはない 1 つの有害事象（Adverse Event）が含まれていた。

複数の有識者に、この結果に対する見解を求めた。ある有識者は、「いわゆる『デバイスラグ』のため、先行販売された欧州・米国において新たな有害事象が発見され、日本の添付文書に書き加えざるを得なかったのではないか」との見解を示した。一方で、「日本と海外とで完全に文書を一致させようとする事業者とそうでない事業者がいる」と指摘する有識者もいた。後者の指摘は、事業者が国民性に応じてリスクの提示方法・内容を書き分けていないとも限らないことを示唆している。これらの書き分けが添付文書と Labeling に現れているとすれば、テキストマイニングによりリスク認知に関する何らかの知見を得られることが期待される。

本研究では、添付文書/Labeling に対するテキストマイニングにより、リスク認知に関する基礎的知見を得ることを目的とした。

1.2 本報告書の構成

本報告書は、4 つの章で構成されており、第 1 章は本章であり、研究の背景及び目的と方針について述べている。

第 2 章では、解析手法について述べる。解析対象とした添付文書/Labeling の収集方法について述べた後、構築した解析プログラムについて述べる。

第 3 章では、解析プログラムによって添付文書/Labeling を解析した結果について述べるものである。

第 4 章では、本研究で得られた知見を要約して述べる。

2　解析手法

2.1 添付文書/Labeling の調査と収集

2.1.1　添付文書

医療機器の添付文書は、独立行政法人医薬品医療機器総合機構（以下、「pmda」という。）のホームページから取得可能である(1)。添付文書は、HTML 形式、DTD 形式、SGM 形式あるいは PDF 形式のいずれかを選択できる。本研究で選択したファイル形式は PDF である。これは、後述するが、Labeling のファイル形式と整合させることで、プログラム上での処理を容易にするためである。

公開されている添付文書数は、16,154 件であった（2012 年 7 月 6 日現在)。ただし、全ての医療機器の添付文書が登録されているわけではない。解析に必要なファイルを適宜ダウンロードしている。添付文書のページ数は、おおよそ 8 ページほどである。また、ファイルサイズはほぼ 1MB 未満である（PDF の場合)。添付文書の構成は、おおよそ禁忌・警告、使用方法、不具合・有害事象、臨床試験成績から成る。

2.1.2　Labeling

医療機器の Labeling は、米国食品医薬品局（Food and Drug Administration, 以下「FDA」という。）のホームページから取得可能である(2)。Labeling は、510(k) あるいは PMA プロセスによって申請された医療機器が公開されている。今回、PMA プロセスの医療機器の Labeling を収集した。PMA プロセスの Labeling は、510(k)と比較して容易に入手可能であるというメリットがある。

PMA プロセスの Labeling は、1994 年頃のものから公開されている。いずれも紙資料をスキャンしたものであり、ファイル形式は PDF である。1994 年から 2000 年以前までの PDF ファイルは、文字が図として認識されている。このため、プログラムから自動的にテキストデータを抽出することができない。そこで、文字としてプログラムから自動的に抽出できる 2000 年以降の全 Labeling を収集した。取得した Labeling 数は 415 ファイルである（2012 年 10 月 30 日現

在)。Labeling の構成は、おおよそ禁忌・注意、臨床試験成績、不具合・有害事象、操作法、患者向け資料から成る。

2.1.3 添付文書/Labeling の比較

上述した通り、添付文書と Labeling との間には、ファイル形式、ページ数、ファイルサイズあるいは構成に違いがみられる。特に、構成において、Labeling では患者向け資料が含まれていることは大きく異なる。このことは、添付文書に比べて Labeling のページ数が多いことに寄与する一因となっている。

表 1 添付文書と Labeling の比較

項目	添付文書	Labeling
公開元	独立行政法人 医薬品医療機器総合機構	U.S. Food and Drug Administration
公開ファイル形式	HTML, DTD, SGM, PDF	PDF
ページ数	約 8 枚	およそ 50 枚以上
ファイルサイズ(PDF)	ほぼ 1MB 未満	ほぼ 1-2MB 程度
公開数	16,154 (2012 年 7 月 6 日現在)	415 (2010 年以降の PMA のみ) (2012 年 10 月 30 日現在)
構成（代表例）	禁忌・警告、使用・操作方法、不具合・有害事象、臨床試験成績 等	禁忌・注意、臨床試験成績、不具合・有害事象、使用・操作法、患者向け資料 等

2.1.4 解析対象

解析対象としたのはステントグラフトの添付文章/Labeling である。ステントグラフトを選んだのは有害事象に係る記載が多く、リスク認知の研究のために適していると考えたためである。

日本と米国で販売されている同じステントグラフトの添付文書/Labeling

をそれぞれ入手して解析を行った。文章数はそれぞれ3つである。

2.2 解析プログラム構築

2.2.1 テキストマイニング

緒言においても述べたが、ここでは改めてテキストマイニングについて紹介する。テキストマイニングにおける解析の対象はテキストである。このテキストとは、定型化されていないデータである。従来、データ解析では、定型化されたデータを取り扱うことが主であった。定型化データとは、例えば気温、湿度等の項目（変数）が、時系列、場所毎に整理された表形式（あるいは）配列形式にまとめたものを指す。定型化されていないテキストに対してデータ解析を行う場合、テキストを定型化する必要がある。そこでテキストを、単語やフレーズ等の単位に分割し、それらの出現頻度や共起関係（同時出現）などを集計して表形式に変換することで定量的に解析することが可能となる[5]。通常、文章を人手により解析することは現実的に難しいため、コンピュータを用いる。特に大規模なデータでなければ並列分散処理する必要がなく、PCでも十分解析が可能である。

定型化されていないテキストを、定型化データに変換することは、テキストマイニングの重要な技術の一つに挙げられる（ベクトル化とも呼ばれる）。テキストが定型化データに変換された後の解析には、一般的な統計解析手法（データマイニングとも呼ばれる）を応用することができる。これらは、例えば、クラスタリング、グラフマイニング、情報抽出、時系列分析予測、関係マイニング、バスケット分析相関ルール、リコメンデーション、主成分分析や異常検知等がある[6]-[10]。テキストマイニングの分野でよく用いられるのは、クラスタリングや主成分分析等である。これらの分析をパッケージとして組み込んだソフトウェアも販売されている。

ところで、テキストマイニングをリスク認知研究に応用することは比較的新しい試みであり、従来の分析手法や既存のソフトウェアが有効かどうかは不明であった。このため、本研究では、公開されているテキストマイニング関連の

API[7]を活用しつつ、リスク認知研究のための独自の解析プログラムを構築する。以下に、開発環境、構築したプログラム、解析に要した医学用語辞書の構築手法について述べる。

2.2.2 解析環境及び言語

PC の環境は、32bit の WindowsXP Professional Version 2002 Service Pack3 であり、CPU は Intel Core i5 3.10GHz である。オープンソースの統合開発環境 Eclipse を用いた。Eclipse のバージョンは Helios である(3)。プログラミング言語は、公開されているテキストマイニング関連の API が豊富であるという理由から Java 言語を選んだ。

2.2.3 PDF 文書からのテキスト抽出手法

添付文書/Labeling はいずれも PDF 形式である。PDF からテキストを抽出するための API を調査・検討した。調査の結果、iText と Apache PDFBox(4)（以下、「PDFBox」という。）が候補に挙がった。これらで同一の日本語の PDF ファイルのテキストを抽出した結果を以下に述べる。

テスト時においては、添付文書ではなく審査報告書を用いた。表 2 の通り、iText で読み取ったテキストには不要な記号が含まれていた。一方、PDFBox ではテキストをほぼ正常に抽出することができた[8]。iText によるテキストの抽出には、さらに多くの設定が必要と考えられる。このため、本研究では PDFBox を採用することとした。

---

[7] Application Programming Interface: API は、アプリケーションから利用できる、オペレーティングシステムやプログラミング言語で用意されたライブラリなどの機能の入り口となるものである。主に、ファイル制御、ウインドウ制御、画像処理、文字制御などのための関数として提供されることが多い。（API. In Wikipedia: *The Free Encyclopedia.*）

[8] 常用漢字以外は抽出できないようである。

表 2 審査報告書のテキストの抽出結果比較

iText	PDFBox
/GS1 gs	審査報告書
BT	平成 21 年 12 月 8 日
/TT2 1Tf	独立行政法人医薬品医療機器総合機構
10.02 0 0 10.02 85.08 51.8604 Tm	
(後略)	*(後略)*

今回、PDFBox を利用して PDF ファイルからテキストを抽出する Java プログラムのメソッドを以下のように定義した。このメソッドで、添付文書/Labeling から日本語/英語をそれぞれ抽出することができる。

実装例 1　PDFBox を利用した PDF からのテキスト抽出メソッド[9]

```
import org.apache.pdfbox.pdmodel.PDDocument;
import org.apache.pdfbox.util.PDFTextStripper;
private String getPDFText() throws IOException{
          PDDocument pdd = PDDocument.load(new File(PDFfilepass));
          PDFTextStripper sr = new PDFTextStripper();
          String str = sr.getText(pdd);
          pdd.close();
          return str;
}
```

2.2.4　解析手法

任意の文章を、意味を持つ最小の文字列の単位（以下、「形態素」という。）に分割し、品詞情報などを付け加えることを形態素解析という。形態素に分割することを、形態素分割という。日本語は、英語と異なり形態素の境界判別のために形態素分割は必須といえる。

[9] 斜字の部分は、環境に合わせてユーザーが指定する変数である。

インターネット上において形態素解析のフリーソフトがいくつか公開されている。これらのフリーソフトの中でも、NTT コミュニケーション科学基礎研究所の共同研究プロジェクトを通じて工藤拓（現 google）が開発した MeCab は有名である。この MeCab は C++で開発されたものであるが、MeCab を Java に移植したものは Sen と呼ばれる(5)。

この Sen を開発環境にインストールし、添付文書に記載されたテキストの一部を形態素分割した結果を表 3 に示す。なお、Sen により形態素分割するメソッドを実装例 2 に示す。

表 3 が示す通り、Sen による形態素分割では、一部の不具合事象を表す単語が形態素に分割されてしまい、元の単語が持つ厳密的な意味を失っている。これは、Sen の形態素解析エンジンで使用している辞書は、医学用語を網羅していないことが主たる原因とみられる。表 3 の例では、「対麻痺」が「対」と「麻痺」に分割されている。Sen には、医学用語の形態素解析のための辞書が公開されているものの(6)、完全には医学用語の境界区分が難しいことが指摘されている。

表 3　Sen による形態素解析事例(分割結果のみ)

テキストデータ	形態素分割結果 (カンマが区切り位置)
局所性、または全身性の神経学的合併症、及び後遺症（脳卒中、対麻痺、不全対麻痺、麻痺等）	局所, 性, 、, または, 全身, 性, の, 神経, 学, 的, 合併症, 、, 及び, 後遺症, (, 脳卒中, 、, 対, 麻痺, 、, 不全, 対, 麻痺, 、, 麻痺, 等,)

実装例 2　Sen による形態素解析

```
import net.java.sen.StringTagger;
import net.java.sen.Token;
public Token[] jaTokenize(String str) throws IllegalArgumentException,
IOException{
        StringTagger tagger = StringTagger.getInstance();
        Token[] token = tagger.analyze(str);
        return token;
}
```

リスク認知を研究しようとする上で、リスクを表す単語である不具合事象が不適切に分割されてしまうことは望ましいといえない。さらに、対応する英単語を同定できない問題が生じる。さきほどの「対麻痺」の例では、対応する英単語は「paraplegia」であるが、「対」「麻痺」では「pair」「paralysis」となる。このままでは、添付文書/Labeling の内容を比較することができない。そこで、有害事象に関しては、医学用語の和英辞書（以下、「医学用辞書」という。）を準備し、この辞書の単語を添付文書/Labeling にマッチングさせることで有害事象を抽出することとした。

2.2.5　医学用辞書の構築

医学用辞書は、インターネット上で公開されている医学用語の和英辞書を利用することとした。様々なデータベースを調査・検討したところ、ライフサイエンス総合データベースと北里大学医療衛生学部 和英医学用語集は、医学用語を広く網羅する辞書であることが分かった。ライフサイエンス総合データベースには、150,592 語（重複あり）の医学用語が登録されている(7)（2013 年 1 月 23 日現在)。また、北里大学医療衛生学部 和英医学用語集には、30,178 語が登録されている(8)（2013 年 1 月 23 日現在)。この 2 つの辞書を組み合わせることで、医学用語を相互に補完させることとした。

ライフサイエンス総合データベースは、1 つのレコード内で日本語と英語は相互に一意である。しかしながら、北里大学医療衛生学部 和英医学用語集は、

英単語に対して複数の日本語が一つのレコード内に登録されている。例えば、以下のような例である。

表 4 北里大学医療衛生学部 和英医学用語集の例

和文	英文
1.破傷風、テタヌス 2.破傷風強直(持続性筋強直) 3.強直(性)けいれん、強直	Tetanus

表 4 に示すように、Tetanus という英単語に対して、「破傷風」、「テタヌス」、「破傷風硬直」、「持続性筋強直」、「硬直性けいれん」、あるいは、「硬直けいれん」、「硬直」という 7 つの単語が対応する。

本研究では、1 つのレコード内において、1 つ英単語に対して複数の日本語が対応する場合に、プログラム上で一意の日本語となるよう 1 つのレコードを複数のレコードへ変換処理している。具体的には、括弧、読点や数字の排除等である。表 4 の例では、処理後に以下のような複数のレコードが作成される。

表 5 北里大学医療衛生学部 和英医学用語集の変換処理後の例

日本語	英単語
破傷風	Tetanus
テタヌス	Tetanus
破傷風強直	Tetanus
持続性筋強直	Tetanus
強直性けいれん	Tetanus
強直けいれん	Tetanus
強直	Tetanus

また、1 つの日本語の中に括弧が複数含まれる場合には、括弧内の単語の有り・無しの場合の全ての組み合わせを取得するよう処理する。例えば、表 6

のような複数のレコードが作成される。当然、組み合わせ次第では、無意味な単語を形成する場合もある。しかしながら、このような単語はマッチングにおいて無視されるため問題ない。

表 6 北里大学医療衛生学部 和英医学用語集の括弧が 2 つ以上の変換例

変換前	肺(動脈)塞栓(症)	pulmonary embolism
変換後	肺塞栓	pulmonary embolism
	肺動脈塞栓	pulmonary embolism
	肺塞栓症	pulmonary embolism
	肺動脈塞栓症	pulmonary embolism

上記のような処理をした医学用辞書の単語を、それぞれ添付文書/Labeling のテキストにマッチングさせて単語を抽出した。

医学用辞書は、医学用語を全て網羅しているため、マッチング抽出単語は有害事象に限らず、治療方法（手術）、器官、臨床検査をはじめとするあらゆる用語が抽出される。これらの単語が全てリスクを表すとも限らない。このため、マッチング抽出した単語を、リスクを表すものとそうでないものに分類する必要がある。

最も単純な方法として、医学用辞書の単語を全て精査して分類することである。しかしながら医学用辞書は、上述の変換後に 20 万語以上になるため、これらを人手によって分類することはあまり現実的ではない。そこで、病名のみから成る辞書を構築することとした。病名を一覧で公開しているデータベースをインターネット上で検索したところ、医学情報研究所が病名一覧を公開していることがわかった(9)。本研究では、このデータベースから新たに辞書（以下、「疾病名辞書」という。）を構築し、疾病名か否かの判別に利用することとした。

2.2.6 集計･分析

現在までに実装している解析プログラムの出力を表 7 にまとめる。また、解析プログラムの出力例を表 8 に示す。今回は、マッチング抽出した単語のみを解析の対象としている。

表 7 集計・分析結果の出力一覧

項目	メソッド	説明
マッチング抽出単語	頻度	添付文書/Labeling に出現した回数を出力する。
	TF-IDF スコア	TF-IDF スコアとは、語の重みの指標である。多くのテキスト（添付文書）に現れる場合にはスコアが小さくなり、特定のテキスト（添付文書）にしか現れない場合には大きくなる。TF-IDF は、テキスト d における語 t の頻度 tf、語 t を含むテキスト数 df、テキスト総数 N とした時、以下の式で表される。 $$\mathrm{TF-IDF} = tf \log\left(\frac{N}{df}\right)$$ 上式で算出した TF-IDF スコアを出力する。
	特異語抽出	添付文書のみ、または、Labeling にのみ現れる医学用語である。特異語である場合には 1 を出力する。
	疾病名抽出	疾病名辞書に現れる用語。現れる場合には 1 を出力する。(特異語かつ疾病名の場合を、以下「特異疾病名」という。)
	訳語	英語の場合は、対応する日本語を全て出力する。
	前後文脈表示	医学用語の前後の文脈を表示する。添付文書は前後 10 語、Labeling は前後 20 語を出力する。

表 8 解析プログラムによる集計・分析結果の出力の一例

種類	医学用語	頻度	TF-IDF	特異語	疾病名	訳語	前後文脈(『』内はマッチング抽出した単語)
Labeling	angina pectoris	3	3.295836866	0	1	[狭心症]	・・・ocumented 『angina pectoris』 or functi・・・・・・ocumented 『angina pectoris』 and a sin・・・・・・ocumented 『angina pectoris』 or functi・・・
Labeling	hemorrhage	3	3.295836866	0	1	[出血]	・・・n such as 『hemorrhage』, embolism・・・・・・included: 『hemorrhage』 (gastric ・・・・・・ecified), 『hemorrhage』 (cathctcr・・・
Labeling	impaired	3	3.295836866	0	1	[障害, 欠陥的]	・・・ents with 『impaired』 renal fun・・・・・ents with 『impaired』 renal fun・・・・・・ents with 『impaired』 renal fun・・・
Labeling	asymmetry	3	3.295836866	0	0	[非相称, 非対称性, 不斉, 左右不同, 不整, 非対称, 無対称]	・・・lved mild 『asymmetry』 or stent ・・・・・・nresolved 『asymmetry』 or stent ・・・・・・rrect the 『asymmetry』. The seco・・・
Labeling	atherectomy	3	3.295836866	0	0	[アテレクトミ, 粥腫切除術]	・・・echanical 『atherectomy』 devices (・・・・・・rectional 『atherectomy』 catheters・・・・・・otational 『atherectomy』 catheters・・・
Labeling	butylene	3	3.295836866	0	0	[ブチレン]	・・・rene-b-iso 『butylene』 -b- styren・・・・・・ne and iso 『butylene』 units bui・・・・・・its of iso 『butylene』 1.2.3 Pro・・・
添付文書	動脈狭窄	2	2.197224577	1	1	null	・・・胸痛（2.0%）、冠『動脈狭窄』（1.7%）、虚血性・・・・・・悪化（3.9%）、冠『動脈狭窄』（2.8%）、心房細・・・
添付文書	紫斑病	2	2.197224577	1	1	null	・・・、血栓性血小板減少性『紫斑病』（TTP）、無顆粒球・・・・・・ら血栓性血小板減少性『紫斑病』、顆粒球減少、肝障害・・・
添付文書	細動	2	2.197224577	1	1	null	・・・動脈瘤 7) 心室『細動』（VF）、心室頻拍（・・・・・・窄（2.8%）、心房『細動』（2.7%）、心室性・・・
添付文書	血小板減少性紫斑病	2	2.197224577	1	1	null	・・・与においては、血栓性『血小板減少性紫斑病』（TTP）、無顆粒球・・・・・・患者の状態から血栓性『血小板減少性紫斑病』、顆粒球減少、肝障害・・・
添付文書	金属アレルギー	2	2.197224577	1	1	null	・・・が溶出することにより『金属アレルギー』を惹起するおそれがあ・・・・・・ので、必ず問診を行い『金属アレルギー』の患者についてはステ・・・
添付文書	BA	2	2.197224577	1	0	null	・・・E、ポリアミド、PE『BA』X、ポリエチレン バ・・・・・・ン バルーン： PE『BA』X ●原理 抗増殖・・・

3　解析結果及び考察

本章では、上述の解析プログラムによって、3 つのステントグラフトの添付文書/Labeling を解析した結果を示す。なお、1 つのファイルを処理するのに必要な時間は 5-10 秒ほどであった。

3.1 出現頻度比較

医学用辞書の単語が、添付文書/Labeling に出現する頻度を、ランク順に並べたものを図 1 に示す。両軸とも対数で示している。医学用辞書にマッチングした単語数は、添付文書で 1,353 語であり、Labeling で 1,534 語であった。Labeling のほうが多いが、これは Labeling の単語数そのものが多いことが起因している。具体的には、3 つの添付文書の文字数の算術平均値は 20,146 語であるが、Labeling は 106,131 語である。単純に比較すれば、Labeling は添付文書よりも 10 万ほど語数が多いことになる。

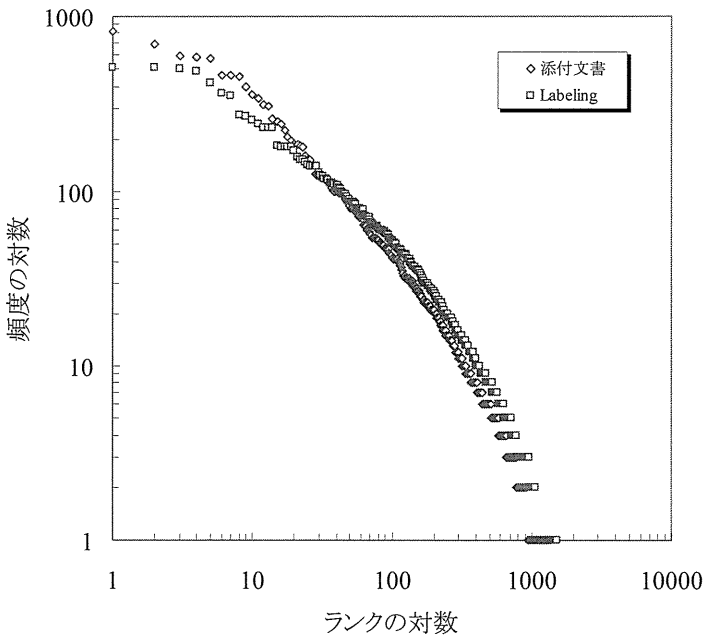

図 1 添付文書/Labeling の出現頻度比較