

医療記録における語彙の月別出現分布

相良かおる（西南女学院大学 保健福祉学部）

山崎 誠（国立国語研究所）

中島直樹・山下貴範（九州大学病院）

小野正子（西南女学院大学 保健福祉学部）

本稿の目的は、医療記録に含まれる語彙の季節変動を調べることにある。今回 12 カ月分の医療記録を対象に各月の医療記録に特徴的に多く出現する名詞を対数尤度比検定により特定し、対象月と参照月で共通する特徴語数を求め季節変動を調べた。その結果、2 月と 3 月、4 月と 5 月というように連続する 2 カ月間で共通する特徴語数が多いことが示唆された。

Monthly variation in vocabulary included in medical records

Kaoru Sagara (Faculty of Health and Welfare, Seinan Jo Gakuin University)

Makoto Yamazaki (Research Dept., NINJAL)

Naoki Nakashima / Takanori Yamashita (Kyushu University Hospital Medical Information Center)

Masako Ono (Faculty of Health and Welfare, Seinan Jo Gakuin University)

The aim of this paper is to examine seasonal variation in vocabulary included in medical records. Nouns appearing characteristically frequently in monthly medical records over a period of 12 months were identified through a log-likelihood ratio test, numbers of characteristic words common to target months and reference months were obtained, and seasonal variations were examined. Results suggested high frequency of numbers of characteristic words common to successive two-month periods, for example, February and March or April and May.

1. はじめに

国民の医療情報を匿名加工し、大学や製薬企業の研究開発などでの活用を可能にする仕組みを定めた「次世代医療基盤法」が、2017 年 4 月 28 日に、国会で可決・成立し、今後、医療情報の 2 次利用が活発化すると推測される。

蓄積された医療情報を 2 次利用するためには、匿名加工処理に加え、倫理審査を含め様々な手続きが必要であり、①何のために（目的）、②どのようなデータが（種類）、③どれだけ（量）必要かを明確にする必要がある。

また年間に蓄積される医療情報は膨大で、これらを機械的に処理するには環境整備も必要となる。そこで、母集団を代表する標本の抽出方法に関する情報は有用である。

筆者等は、医療情報の自然言語処理を支援するために 2004 年より実践医療用語の収集・分析を始め、2008 年より分ち書き用の辞書 ComeJisyo¹⁾（登録語数 30,146 語）を作成し、以降随時更新し、2013 年には ComeJisyoV5-1（登録語数 77,760 語）を公開している。

実践医療用語の収集は、医療施設において倫理審査等の諸手続きおよび匿名加工処理を施した 1 年分の看護経過記録、プログレスノート、医師診療記録等の医療情報から機械的に専門用語候

補を抽出し、選定作業は人手で行っている。しかしながら、これらは、量と質の両面において大変な作業量であり、効率的なデータの採取法は検討課題である。

小児科、歯科、外科等の診療の専門分野によって医療情報に含まれる専門用語に相違があることは明らかであり、また看護師、助産師、管理栄養士の国家試験に含まれる語彙調査の結果、同じ用語においても、職種によって教育段階で「癌」と「がん」、「Alzheimer 型認知症」と「アルツハイマー型認知症」等の表記のゆれがあることが分かっている²⁾³⁾。一方、医療情報に含まれる語彙の季節変動についての調査研究は見当たらない。

そこで、効率の良い標本抽出法について、夏と冬の半年のデータから無作為にデータを抽出すれば良いのか、春夏秋冬それぞれ 3 か月分のデータから無作為にデータを抽出すればよいのか、各月毎のデータが必要であるのか等の根拠を得るために、1 年分の医療記録において、月別データから特徴語を機械的に選定し、4 月と 5 月というように ${}_{12}C_2$ の 66 通りについて共通する特徴語数を求め語彙の月別変動の調査を行った。本稿では、これらについて述べる。

2. 関連研究

医療記録に限らず、テキストに含まれる語彙または文字列の季節変動に関する調査研究は見当たらない。しかしながら、「季節変動を調べる」を、季節毎にまとめられたテキストに含まれる特徴的な語または文字列の差異を調べることを考えると、統計的尺度を用いて複数のコーパスから特徴語や重要語を特定する研究、また、英語教育のための分野特徴単語の選定尺度の比較等の報告がある⁴⁾⁵⁾。

そして人手による語彙の選定は、コストが高い上に質的な面で選定者の主観や経験に依存することから、大量のテキストデータから特徴的な語彙を機械的に選定する統計的尺度として、①相互情報量、②カイ2乗値、③対数尤度比、④tf・idf等が提案されている⁶⁾。

3. 調査方法

3.1 調査データ

本調査では、諸手続きおよび匿名加工処理済みの1施設1年分の医師診療記録データ、3,107,547行(596MB)を調査データとする。各月別の行数は表1に示す通りである。

表1 月別医療記録の行数		
	行数	割合
4月	273,216	8.8%
5月	268,480	8.6%
6月	264,172	8.5%
7月	283,492	9.1%
8月	271,120	8.7%
9月	249,109	8.0%
10月	262,918	8.5%
11月	231,171	7.4%
12月	253,141	8.1%
1月	240,581	7.7%
2月	239,884	7.7%
3月	270,263	8.7%
合計	3,107,547	100.0%

なお、ここでの1行は1文ではなくデータベース上の1レコードを意味し、1レコードの中には、複数の文が含まれる場合が多い。例えば4月のデータの中には、1行に380文(文字数9,056文字)、から1文1文字の行迄ある。

従って、各月のデータの文字数のばらつきは大きく、均等なものではない。

以下に記録の1行(6文351文字)の例を示す。

XXXX年4月中旬くらいから食後の腹痛が出現し、排便回数が増加した(3回/日)。5月くらいから便に血が混じるようになった。排便回数5-6回/日と増加し、症状の増悪があったため、5月13日にXXX科医院を受診された。採血では、Hb 5.7g/dlの貧血があり(CRP 未)、浣腸の前処置でS状結腸内視鏡検査を行ったところ、下行結腸から直腸にかけて連続性に血管透視不良、びらん、浮腫状粘膜を認め、潰瘍性大腸炎を疑う所見であった。生検でも陰窩膿瘍等の所見あり、潰瘍性大腸炎に矛盾しない所見であり、5月20日、当科紹介。以後、潰瘍性大腸炎(全大腸型)として外来加療中(5-ASA、PSL、AZA内服)であったが、PSL漸減に伴い排便回数・血便などの症状増悪あり、ヒュミラ導入目的にXXXX年3月31日、入院となった。

図1は、上記の文を形態素解析器 MeCab0.996⁷⁾と医療実践用語辞書 ComeJisyoV5-1¹⁾をユーザ辞書として使って分かち書きした結果である。

本調査データは、この様に語分割したデータに、3.2節に記載の前処理を実施したものである。

3.2 前処理

形態素解析器を用いた語分割には、誤解析が含まれる(図1)。また、本調査では特徴語として事物や概念を表す名詞のみを対象とする。

そこで、語分割した213,758語(延べ語数128,039,037語)に対し、以下の前処理を行う。

- (1) 数字および記号のみの語を削除
- (2) 品詞が名詞のもの、および品詞が感動詞となっている英語(アルファベットからなる文字列)を抽出
- (3) 各月別に(2)の語の出現頻度を求める
- (4) 各月の頻度が12以下の語について、目視にて以下のもの(解析あやまり)を削除
 - ・ 小文字のひらがなまたはカタカナから始まるもの
 - ・ “ヲ”または“ン”で始まるもの
 - ・ 文字長が1文字の語
 - ・ 文字長が20文字以上のひらがな
- (5) 各月の頻度が全て1または0の語を削除
- (6) 各月別に語頻度の平均と中央値を求める
- (7) 全ての月において頻度が各月の平均以上の語を削除

上記の前処理を実施した後の異なり語78,762語を本調査データとする。表2は前処理後の語の概要である。

平均値と中央値の開きが大きいことから正規性がないことは明らかである。

XXXX | 年 | 4 | 月 | 中旬 | くらい | から | 食後 | の | 腹痛 | が | 出現 | し | 、 | **排便回数** | が | 増加 | し | た | (| 3 | 回 | / | 日 |) | 。 | 5 | 月 | くらい | から | 便 | に | 血 | が | 混じる | よう | に | なっ | た | 。 | 排便回数 | 5 | - | 6 | 回 | / | 日 | と | 増加 | し | 、 | 症状 | の | **増悪** | が | あっ | た | ため | 、 | 5 | 月 | 13 | 日 | に | 大 / xxx / 科医 / 院 | を | 受診 | さ | れ | た | 。 | 採血 | で | は | 、 | **Hb** | 5 | . | 7 | g / / / dl | の | 貧血 | が | あり | (| **CRP** | 未 |) | 、 | 浣腸 | の | **前処置** | で | S / 状結腸 | **内視鏡検査** | を | 行っ | た | ところ | 、 | **下行結腸** | から | 直腸 | にかけて | **連続性** | に | 血管 | 透 | 見 | 不良 | 、 | **びらん** | 、 | 浮腫状 / 粘膜 | を | 認め | 、 | **潰瘍性大腸炎** | を | 疑う | 所見 | で | あっ | た | 。 | **生検** | で | も | 陰 / 窩 / 膿瘍 | 等 | の | 所見 | あり | 、 | **潰瘍性大腸炎** | に | 矛盾 | し | ない | 所見 | で | あり | 、 | 5 | 月 | 20 | 日 | 、 | 当科 | 紹介 | 。 | 以後 | 、 | **潰瘍性大腸炎** | (| 全 / 大腸 / 型 |) | として | 外来 / 加療 / 中 | (| 5 / - / AS / A | 、 | **PSL** | 、 | **AZA** | 内服 |) | で | あっ | た | が | 、 | **PSL** | 漸減 | に | 伴い | **排便回数** | ・ | 血便 | など | の | **症状増悪** | あり | 、 | **ヒュミラ** | 導入 | 目的 | に | XXXX | 年 | 3 | 月 | 31 | 日 | 、 | 入院 | と |

図 1 形態素解析の結果

注：ComeJisyoV5-1に登録されている語を太字に、登録されておらず、過分割された語を下線付きの斜体にしている。

表 2 対象データ (語) の概要

	異なり語数	割合	平均値	中央値	延べ語数	割合	平均値	中央値	最大値	最小値
4 月	51,511	65.4%	77	5	3,945,176	9.3%	77	5	83,172	1
5 月	52,054	66.1%	77	5	4,016,875	9.5%	77	5	82,565	1
6 月	51,683	65.6%	76	5	3,931,218	9.3%	76	5	79,781	1
7 月	50,338	63.9%	70	5	3,513,082	8.3%	70	5	76,028	1
8 月	49,371	62.7%	68	5	3,365,410	8.0%	68	5	75,212	1
9 月	49,444	62.8%	67	5	3,302,688	7.8%	67	5	72,650	1
10 月	50,444	64.0%	69	5	3,495,654	8.3%	69	5	76,878	1
11 月	48,492	61.6%	63	5	3,069,677	7.3%	63	5	67,833	1
12 月	49,798	63.2%	68	5	3,367,526	8.0%	68	5	73,986	1
1 月	49,757	63.2%	66	5	3,300,926	7.8%	66	5	72,372	1
2 月	49,555	62.9%	67	5	3,300,062	7.8%	67	5	71,284	1
3 月	51,134	64.9%	73	5	3,718,390	8.8%	73	5	81,411	1
合計	78,762	100.0%	840	60	42,326,684	100.0%	537	22	913,172	2

3.3 特徴語の抽出

前述 2 章の 4 種の尺度の内、①相互情報量、②カイ 2 乗値、③対数尤度比は、2 つの確率変数間の依存性の程度 (結合度) を測定するものであり、文書中での語の出現確率等の数値が必要となる。

対して④tf・idf は、語頻度に対数文書頻度の逆数を掛けた値であり、語の出現頻度と文書頻度から容易に計算することが出来る。しかしながら tf・idf は、与えられた文書集合が比較的均一であることが暗黙の前提になっている⁸⁾。

よって本調査で扱う月別データ量は均等ではないことから(表1), ④tf・idfは除外した。

今回, 表3に示す対象月と, 対象月以外の11ヵ月における対象語の語頻度とその他の語頻度の 2×2 の分割表を作成し, 有意差検定により特徴語を特定することとした。

分割表による有意差検定には, ①カイ2乗検定, ②フィッシャーの精密検定, ③対数尤度比検定がある。①カイ2乗検定には, 2つの標本の度数の合計が20以上と十分に大きく, 分割表における全てのセルの度数が4以上という制約が, ②フィッシャーの精密検定には, 一般に観測値の数が少ない, または分割数のセルのいずれかの期待値が4以下に適用されるという制約がある。対して, 対数尤度比検定は, これらの厳しい制約がなく, (1)正規分布を前提としないこと, (2)稀な事象を過剰評価しないこと, (3)標本の分量差が結果に影響しないとされている⁹⁾¹⁰⁾¹¹⁾。

表3 計算のための分割表

	当該月	当該月以外	計
語 W	a	b	a+b
W 以外の語	c	d	c+d
計	a+c	b+d	a+b+c+d

本対象データは, 3.2節の前処理(5)により, 各月の全ての頻度が1または0の語を削除しているが, 各月の頻度の最小値は1語であり, ①カイ2乗検定の制約を満たすことができない。また, 最大値は67,833~83,172語と大きく, ②フィッシャーの精密検定の制約もまた満たしていない(表2)。従って, 本調査では, 対数尤度比を用いて検定統計量を求め, 各月に有意に特徴的な語を特定する。次に, 4月と5月というように ${}_{12}C_2$ の66通りについて共通する特徴語数を求め月別の変動を調べる。

検定統計量は表3のパラメータを用い, 以下の式で求める。

$$\begin{aligned} \text{検定統計量} = & 2 \times [a \times \log(a) + b \times \log(b) + c \\ & \times \log(c) + d \times \log(d) - (a+b) \\ & \times \log(a+b) - (a+c) \\ & \times \log(a+c) - (b+d) \\ & \times \log(b+d) - (c+d) \\ & \times \log(c+d) + (a+b+c+d) \\ & \times \log(a+b+c+d)] \end{aligned}$$

なお, この式で用いるlogは底をeとする自然対数である。また, セルの度数が0, すなわち語頻度が0の場合, log 0は計算不能となるが, その場合係数0と掛け合わせているため, 項全体が0となる。そこで, log 0を0として計算する。

この検定統計量は, ある語の出現確率が対象月とそれ以外の月とで異なる割合を測定することはできるが, 高確率となる場合と低確率となる場合の区別はできない。

そこで, 語 W が高確率で出現するとは, $\frac{a}{c} > \frac{b}{d}$ となることであり, $a \times d - b \times c > 0$ となることであり, 語 W が低確率で出現するとは, ② $a \times d - b \times c < 0$ となることであることから, ②式が成り立つ場合, “-1”を掛けて, 負の数に補正する⁵⁾。

対数尤度検定統計量は, 観測データ数が十分大きい場合, 自由度 k のカイ2乗分布に従うことが分かっている¹²⁾。表4は, 自由度1のカイ2乗分布の有意水準とその臨界値である。本調査では有意水準 $\alpha=0.1\%$ を用いる。

表4 有意水準と臨界値

有意水準	5%	1%	0.5%	0.1%
限界値	3.84	6.63	7.88	10.83

本調査の具体的な手順を以下に示す。

- (1) 各月ごとに語の検定統計量を求める
- (2) 有意に低確率のものと高確率のものを区別するために, 低確率のものがマイナスの値になるように補正
- (3) 有意水準 $\alpha=0.1\%$ として有意に高確率のものと低確率の特徴語を抽出。
- (4) 当該月とそれぞれの月間(${}_{12}C_2=66$ 通り)で共通する特徴語数を求める。

4. 結果

表5 低確率および高確率の特徴語数

$\alpha=0.01\%$	低確率語	高確率語	計
4月	2,899	3,961	6,860
5月	2,760	3,611	6,371
6月	2,522	3,442	5,964
7月	1,929	2,963	4,892
8月	1,961	3,278	5,239
9月	1,828	2,888	4,716
10月	1,849	2,884	4,733
11月	1,682	2,800	4,482
12月	1,859	3,104	4,963
1月	1,977	3,057	5,034
2月	1,949	3,124	5,073
3月	2,408	3,515	5,923

前述の手順で求めた対象語彙データは, 異なり語数 78,762 語, 延べ語数 42,326,684 語であり(表

2), 有意水準を $\alpha=0.1\%$ とし, 帰無仮説を棄却する確率をかなり小さくしたにも関わらず, 表 5 に示す通り, 各月に特徴的な語 4,482~6,860 語が抽出された。

表 6 は 4 月と 5 月, 4 月と 6 月というように 12 カ月から 2 カ月をとる $_{12}C_2$ の 66 通りについて共通する特徴語数をまとめたものである。対角線より右上は, 有意に高確率の特徴語数であり, 対角線より左下は有意に低確率の特徴語である。

図 2 は全特徴語数 (高確率+低確率) をグラフにまとめたものである。5 月の特徴語と共通する特徴語が最も多いのは 4 月, 6 月の特徴語と最も共通する特徴語が多いのは 5 月である。同様に, 7 月では 6 月が, 8 月では 7 月が, というように連続する月において共通する特徴語数が多くなっている。また, 4 月 5 月 6 月の 3 か月間と, 1 月 2 月 3 月の 3 か月間において共通する特徴語数が多くなっている。

表 6 共通する特徴語の数

	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月	1 月	2 月	3 月
4 月	3,961 2,899	907	595	194	155	102	85	80	75	88	80	108
5 月	1239	3,611 2,760	786	193	140	89	72	81	67	67	56	68
6 月	957	1118	3,442 2,522	454	153	99	77	62	64	73	53	85
7 月	540	561	638	2,963 1,929	463	193	111	86	118	110	89	106
8 月	452	431	440	661	3,278 1,961	528	187	125	155	118	108	132
9 月	351	335	345	497	656	2,888 1,828	385	151	158	109	101	134
10 月	342	282	285	400	454	604	2,884 1,849	439	217	137	133	137
11 月	277	250	220	306	354	450	609	2,800 1,682	473	217	157	183
12 月	284	239	228	304	373	438	509	622	3,104 1,859	490	236	256
1 月	346	264	201	289	331	414	476	523	688	3,057 1,977	564	275
2 月	341	270	232	282	311	375	440	463	572	818	3,124 1,949	694
3 月	494	379	327	352	352	474	503	501	593	767	947	3,515 2,899
	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月	1 月	2 月	3 月

※対角線より右上：有意に高確率（限界値 10.83 以上） 対角線より左下：有意に低確率（-10.83 以下）

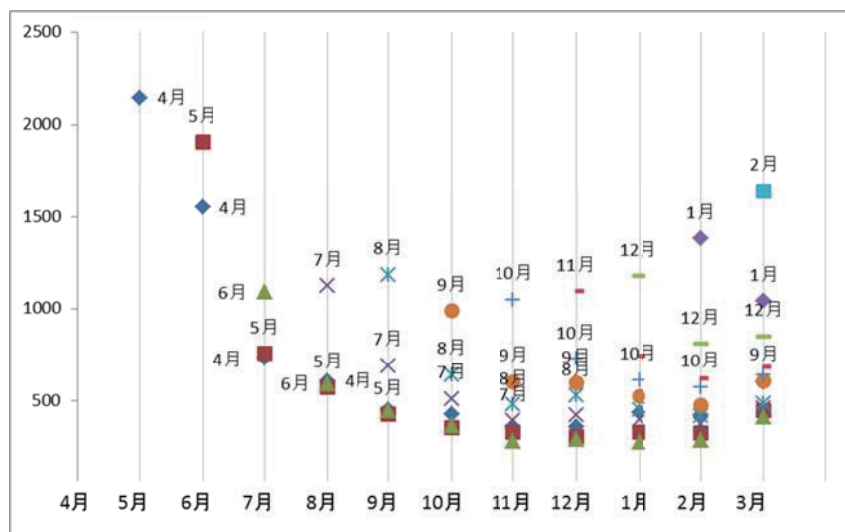


図 2 共通する特徴語数

表 7 高確率の特徴語上位 10 位 (英語や略語等の英字文字列を除く)

	1 位	2 位	3 位	4 位	5 位	6 位	7 位	8 位	9 位	10 位
4 月	造影剤	記入	問診	副作用	アレルギー疾患	体質	ヨード	医師	腎機能障害	有無
5 月	造影剤	記入	ヨード	副作用	アレルギー疾患	体質	連休	ペースメーカー	腎機能障害	問診
6 月	造影剤	記入	ヨード	体質	アレルギー疾患	副作用	ペースメーカー	6 月	問診	有無
7 月	フラジール	ハイカリック	アルブミン	7 月	カルデナリン	輸液	アミロイド	台風	8 月	幻肢痛
8 月	日齢	お盆	夏休み	年生	プレアミン	光線療法	スターシス	新生児低血糖	8 月	びまん性
9 月	ダブラフェニブ	病的肥満症	9 月	オキノーム	尺骨遠位端骨折	献血	ボディー	麻痺性イレウス	若年性関節リウマチ	全身性アミロイドーシス
10 月	予後因子	平均圧	ノボヘパ	メデット	輸液	ノボラピット	オブジーボ	マインドフルネス	身だしなみ	新生児呼吸障害
11 月	小計	オブジーボ	ポンピング	リフレ	大動脈狭窄	豹紋状	周辺	きよくかきん 棘下筋	肩甲下筋	ケラト
12 月	年内	年末年始	クリスマス	来年	年末	プロタノール	ブスルファン	バ・元・ット	年始	母乳
1 月	正月	インフルエンザ	年末	年末年始	お正月	らせん状	上方	風邪	バ・元・ット	色素性
2 月	インフルエンザ	内皮	イレウス管	上方	セロトニン症候群	黄土	分後	ジェイズロフト	ポケット	虹彩新生血管
2 月	C e t	インフルエンザ	Stool	内皮	イレウス管	kgH	Tos	Spo	EES	Esophagectomy
3 月	アトシ	末梢神経障害持続	ジーラスタ	ストラテラ	乗車	支持	日目	大腿腫瘍	プロチアデン	急性根尖性歯周炎

注: 斜体 (2 月) は, 英語・略語を含む上位 10 位である。

表 7 は, 各月の英語および略語等のアルファベットのみからなる文字列を除いた高確率の特徴語, 上位 10 位の一覧である。

4 月, 5 月, 6 月には「アレルギー疾患」が高確率に, 1 月, 2 月には「インフルエンザ」が高確率に出現していることが分かる。

そして「連休」「お盆」「夏休み」「年末年始」「風邪」「インフルエンザ」など, 季節性のある語がある一方, そうではない語「造影剤」「記入」「問診」等もある。

表 8 は文字列「アレルギー」を含む特徴語 46 語の内, 有意に高確率 (○印) または低確率 (×印) の 18 語である。「食物アレルギー」は, 通年を通しての疾患であると思われるが, 4 月, 5 月, 6 月に有意に高確率で出現し, その他の月は有意に低確率となっている。表 9 は文字列「インフル」を含む特徴語 13 語の内, 有意に高確率または低確率の 10 語である。春から夏に有意に高確率, 秋から冬には有意に低確率となっている。

5. 考察

各月の医師診療記録データに含まれる名詞を対象に当該月に特徴的な語を特定し, 他の月との比較を行った結果, 図 2, 表 6 に示すように, 連

続する 2 ヶ月間で共通する特徴語数が多くなっていた。

今回, 有意水準 $\alpha = 0.1\%$ と, 帰無仮説 (H_0 : 当該月とその他の月で有意差がない) を棄却し難い設定とする中, 各月で有意に特徴的な語には, 「風邪」「インフルエンザ」など季節性のあるものに加え, 「造影剤」「ペースメーカー」等, 季節性があるとは考え難い語も含まれていた。

これらの季節性がありそうもない語が有意な特徴語として挙げられる, すなわち医師診療記録の記載内容に大きな変化が生じる要因として, 以下のことが考えられる。

- ① 医師の入れ替わり (4 月, 10 月)
- ② 診療報酬制度等の医療に関する制度の変更
- ③ 医学的な発見や副作用等の大きなニュースの発表
- ④ 医療監視の実施
- ⑤ 院内カルテ委員会等の指導
- ⑥ 電子カルテシステムの変更

個々の特徴語に目を転じると, 表 7 の 4 月, 5 月, 6 月の「ヨード」は ComeJisyoV5-1 に未登録の「ヨード治療」「ヨードシンチ」「ヨード造影

剤」,「ヨード造影剤アレルギー」,「ヨードうがい薬」が過分割されたものであるが, ComeJisyoV5-1 の登録語である「ヨードアレルギー」,「ヨードホルムガーゼ」,「ヨード制限」,「ヨード制限食」は過分割されずに一語の特徴語として抽出されている。「ヨードホルム」は ComeJisyoV5-1 には登録されていないが, 未知語と判定され, 品詞「名詞, 一般」として一語となっている。

8月の「^{びまん}瀰漫」は,「瀰漫性」が過分割されたものであり, 2月の特徴語の「黄土」は,「黄土色」が過分割されたもので共に ComeJisyoV5-1 に未登録の未知語である。なお, 2月の医師診療記録データでは, 皮膚の色, および便の色の表現に「黄土色」が頻出している。

表 8 文字列「アレルギー」を含む特徴語

○:有意に高確率 ×:有意に低確率	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月
アレルギー疾患	○	○	○	×	×	×	×	×	×	×	×	×
食物アレルギー	○	○	○	×	×	×	×	×	×	×	×	×
アレルギー	○	○										
アレルギー性肉芽腫性血管炎	○											
アレルギー性血管炎		○										
アレルギー		×							○			
抗アレルギー剤			○									
ヨードアレルギー				○	○							
アレルギー症状				○								
アレルギー反応				×					○			
アレルギー性鼻炎					×							
アレルギー性紫斑病											×	
薬剤アレルギー							○					
金属アレルギー							×					
アレルギー検査								○				
ミルクアレルギー											○	
アレルギー性気管支肺アスペルギルス症												○

表 9 文字列「インフル」を含む特徴語

○:有意に高確率 ×:有意に低確率	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月
インフル		×	×	×	×	×	×		○	○	○	
インフルエンザ		×	×	×	×	×	×	×	○	○	○	
インフルエンザA型										○		
インフルエンザウイルス									○			
インフルエンザチェック										○		
インフルエンザワクチン	×	×	×	×	×	×		○	○			×
インフルエンザ肺炎										○		
インフルエンザ予防接種	×	×		×	×		○	○	○		×	×
インフルエンザ様症状									○			
パラインフルエンザ						○						

表7の2月の斜体の行は, 英語, 略語を含めた全ての語において有意に高確率の上位10位の一覧である。「Spo」は「Spo2」が過分割されたものであり,「Esophagectomy」は「VATS-Esophagectomy」が過分割されたものである。「Spo2」は ComeJisyoV5-1 に登録されているが「Esophagectomy」,そして,「VATS-Esophagectomy」は登録されていない。

形態素解析器 MeCab0.996 は辞書に登録されていない未知語に対し, 品詞列のテーブルにより, 品詞を自動推定する。未知語の品詞は, 自動推定された生起スコアが高い場合は,「名詞, 一般」

となり, 次いで「名詞, 固有名詞」となる。そして最も低い場合は,「感嘆詞」となる⁷⁾。

「Spo2」の一部分である「Spo」のように, ComeJisyoV5-1 の登録語の一部が未知語と判定され, その推定コストが, 複合語の設定コストよりも高くなると過分割され, 品詞は「名詞, 一般」となる。

また,「Esophagectomy」のように ComeJisyoV5-1 に登録されていない語であっても連続するアルファベットや平仮名, 漢字が, 未知語と判定され, 生起コストが高く推定されると「名詞, 一般」として語分割される。

このように、語分割後に前処理を施した対象データ（表 2）の中には、システム辞書およびユーザ辞書（ComeJisyoV5-1）に未登録の未知語と、これらの辞書に登録されている語の一部分が含まれている。

7. まとめ

本稿では、医療記録データの標本抽出法を知るために 1 年分の医師診療記録データを形態素解析器 MeCab0.996 と ComeJisyoV5-1 を用いて語分割し、月別に出現頻度を求め、対数尤度比検定により有意に特徴的な語を特定し、対象月と参照月間に共通の特徴語数を調べた結果について述べた。

医療記録データをコンピュータで解析する際、最初に行われる処理は語分割と品詞の同定である。ところが、MeCab の単語分割・品詞推定の精度（F 値）は 98%とされる¹³⁾。従って、今回の調査データである 1 年分の医師診療記録データ 3,107,547 行を語分割した 128,039,037 語（異なり語 213,758 語）の中の解析誤りは 2,560,781 語程度と推測される。

その上、図 1 のように ComeJisyoV5-1 に未登録の語や、登録語（複合語）の一部分が未知語と判定され、そのコストが高く推定されて、過分割される場合もある。

MeCab を含め、公開されている形態素解析器の解析精度を 100%にすることも、特定領域の用語を網羅した形態素解析用のユーザ辞書（本調査では ComeJisyoV5-1）を作成することも現実的ではなく、調査データが膨大になればなるほど、解析誤りの検出に要する労力も増大する。また、小説や学術論文等とは異なり、医療記録データには誤字・脱字も含まれている。従って、計量的な調査を行う際にも、全数調査よりも母集団を代表する標本調査の方が効率的である。

よって、有意に特徴的な語の中には、明らかに季節性のあるものと、そうではないものが混在しているものの（表 7）、連続する 2 か月間で共通する特徴語が多いことから（図 2、表 6）、隔月の記録データからの標本抽出が適当であると考えられる。

ところで、100%の精度で語分割することが困難な状況において、医療記録データに含まれる語彙の季節変動および周期性を明らかにするためには、予め変動が期待される用語を定め、複数年のデータを対象にした調査が必要であり、さらにその結果の普遍性を検証するためには、複数の医療機関で蓄積された医療記録データによる調査が必要である。

しかしながら、変動が期待される実践医療用語

を洗い出すことは容易ではない。今回の調査により得られた表 5 または表 6 の特徴語には、複合語の一部分が含まれており、これらを利用することで季節変動が期待される用語の洗い出しが可能となる。

なお、筆者らは今回抽出した特徴語を利用し、医師診療記録に使われる実践医療用語の選定を行い、ComeJisyoV6 を作成する予定である。

参考文献

- 1) ComeJisyo:
<http://sourceforge.jp/projects/comedic/> (2017 年 9 月 11 日参照)
- 2) 相良かおる, 小野正子, 上野恵子: 看護師・助産師国家試験に含まれる語彙の調査, 第 17 回日本医療情報学会看護学術大会論文集, 2016.
- 3) 相良かおる, 小野正子, 石井愛子: 看護師・管理栄養士国家試験に含まれる名詞連続語の特徴, 西南女学院大学紀要 Vol.21, p.135-143, 2017.
- 4) 高見敏子: 「高級紙語」と「大衆紙語」の corpus-driven な特定法, 北海道大学大学院国際広報メディア研究科 言語文化部 紀要 Vol.44 p.73-105, 2003.
- 5) 内山将夫, 中條清美, 山本英子, 井佐原均: 英語教育のための分野特徴単語の選定尺度の比較, 自然言語処理 vol.11 No.31, p.165-197, 2004.
- 6) 言語処理学会編: 言語処理辞典, p.350-353, 共立出版, 2009.
- 7) MeCab: <http://taku910.github.io/mecab/> (2017 年 11 月 5 日参照)
- 8) 相澤彰子: 語と文書の共起に基づく特徴度の数量的表現について, 情報処理学会論文誌 Vol.41 No.12 p.3332-3343, 2000.
- 9) G・K・カンジ著, 池谷裕二, 久我菜穂子訳: 「逆」引き統計学, p.88-96, 2009.
- 10) Pascual Cantos Gomez: Statistical Methods in Language and Linguistic Research, p.209-210, Equinox Publishing, 2013.
- 11) 石川慎一郎: 英語コーパスと言語教育—データとしてのテキスト, p.97-102, 大修館書店, 2008.
- 12) 森棟公夫, 照井伸彦, 中川満, 西埜晴久, 黒住英司: 改訂版 統計学, p.342-344, 2015.
- 13) 森田一, 黒橋禎夫: RNN 言語モデルを用いた日本語形態素解析の実用化, 情報処理学会第 78 回全国大会, p.2-13—p.2-14, 2016.