ポスター | 薬剤情報システム

ポスター12
薬剤情報システム

2019年11月24日(日) 10:00 ～ 11:00 ポスター会場1 (国際展示場 展示ホール8)

[4-P1-2-03] 承認医薬品名における形態素構成の調査

○蝦名 勇樹[1]、佐藤 弘康[1]、真井 雄規[1]、島津 智行[1]、喜多 力[1]、田村 広志[1]、渡辺 浩明[1] （1. JA北海道厚生連　帯広厚生病院）

キーワード：Drug Name, Morphological Analysis, Natural Language Processing (NLP)

【背景】医療記録等のテキスト情報には有用な情報が含まれており、自然言語処理 NLP等を用いて新知見を探索する研究が多く行われている。医療テキスト情報の中には医薬品名も存在するが、その活用は十分でない。医療テキスト情報中の医薬品名に関する情報から医薬品をどの程度特定可能であるかを検証するための基礎的検討として、医薬品名を医療薬学上の形態素へ分解するアルゴリズムを用いて、承認医薬品における医薬品名の形態素構成を調査したので報告する。

【方法】2018年12月9日に診療情報提供サービスのホームページよりダウンロードした承認医薬品リストを用いて、その医薬品名を我々が開発したアルゴリズムにより、「一般名」、「ブランド名」、「規格量」、「濃度」、「剤型」、「屋号」等に形態素分解した。解析結果において１つの「ブランド名」や「一般名」が指し示す製品品目数を調査した。また、「規格量」や「剤型」等が記述されている製品の割合についても調査を行った。

【結果】「ブランド名」が検出された医薬品は9,507品目（3,116種類）であった。「ブランド名」のみで一意に製品を特定できる医薬品は1,348品目（43.3%）であった。一方、「一般名」が検出された医薬品は6,930品目（1,149種類）であった。「ブランド名」が検出された医薬品において、「規格量」、「濃度」、「剤型」が検出された医薬品の割合は、それぞれ71.2%、15.7%、94.8%であった。「一般名」は検出された医薬品において、「屋号」が検出された医薬品の割合は69.2%であった。

【考察】本調査より、医療テキスト情報にブランド名や一般名のみ記述されている場合でも、一定の割合で一意に製品を特定できる可能性が示された。本調査結果を詳細に分析することで、医薬品名に関する限られた形態素情報から医薬品を特定するためのマスタを構築可能と考える。

©一般社団法人 日本医療情報学会

承認医薬品名における形態素構成の調査

蝦名勇樹、佐藤弘康、真井雄規、島津智行、喜多力、田村広志、渡辺浩明
JA 北海道厚生連　帯広厚生病院

Investigation of Morphological Composition in Approval Drug Name

Yuki Ebina , Hiroyasu Sato , Yuki Sanai, Tomoyuki Shimazu, Chikara Kita, Hiroshi Tamura, Hiroaki Watanabe
Pharmaceutical Department, Obihiro Kosei General Hospital

In recent years, research on natural language processing (NLP) using text information such as medical records have been actively conducted. In medical text information, drug name often exists, but its use is not sufficient. In order to create a dictionary for identifying drug names from medical text information, we investigated the morphological composition of approved drug names.

“DrugMA” that we developed was used as an algorithm for decomposing drug names into medical pharmacological morphemes. By using DrugMA, the approved drug name was broken down into “general name”, “brand name”, “drug strength”, “drug concentration”, “drug form”, “maker name” and others.

Of drugs with “brand name”, 43.3% were uniquely identifiable only by “brand name”. The most common morpheme pattern for drugs with “brand name” is “brand name” + “drug strength” + “drug form” (4882 drugs, 51.4%).36.6% of drugs with “general name” were uniquely identifiable only by “general name”. The most common morpheme pattern for drugs with “general name” is “general name” + “drug strength” + “drug form” + “maker name” (4058 drugs, 48.1%). By adding coexisting morphemes, the drug identification rate increased.

By creating a dictionary of drug codes in all patterns of coexisting morphemes, it will be possible to specify the drug from the drug name information included in the medical text information.

Keywords: Drug Name, Morphological Analysis, Natural Language Processing (NLP).

1. 背景

医療記録等のテキスト情報には有用な情報が含まれており、自然言語処理 NLP 等を用いた新知見を探索する研究が多く行われている。医療テキスト情報の中には医薬品名も存在するが、その活用は十分でない。医療テキスト情報中の医薬品名に関する情報から医薬品をどの程度特定可能であるかを検証するための基礎的検討として、医薬品名を医療薬学上の形態素へ分解するアルゴリズムを用いて、承認医薬品における医薬品名の構成要素を調査したので報告する。

2. 方法

2018 年 12 月 9 日に診療情報提供サービスのホームページよりダウンロードした約 2 万件の承認医薬品リストを用いて、その医薬品名を「一般名」、「ブランド名」、「規格量」、「濃度」、「剤型」、「屋号」等の医療薬学上の形態素へ分解した。本形態素解析には、我々が Visual Basic for Application にて構築し、Excel のユーザー定義関数「DrugMA」として実装したアルゴリズムを使用した。解析結果において1つの「ブランド名」や「一般名」が指し示す製品品目数を調査した。また、「ブランド名」および「一般名」を基幹語として、それぞれにおける各形態素の組み合わせの割合を調査した。また、各基幹語において、形態素をどこまで指定すると「薬価基準収載医薬品コード」（以下、医薬品コード）が一意に決定できるかを調査するために、医薬品コードの各桁が一意で決まる場合、その数字またはアルファベットとし、一意に決まらない場合、数字の場合は「X」、アルファベットの場合は「@」をワイルドカードとして代用した代表医薬品コードの付与を試みた。

3. 結果

「ブランド名」が検出された医薬品は 9,507 品目（3,116 種類）であった。1,348 品目（43.3%）は、「ブランド名」のみで一意に製品を特定できた（図 1）。また、1つの「ブランド名」が対応する製品数が最も多かったのは、「アムロジピン」であった（表 1）。多くの製品数を有する基幹語であっても、「オルメサルタン」のように代表医薬品コードの先頭 8 桁まで特定可能な医薬品もあれば、「レボフロキサシン」のように代表医薬品コードのすべての桁が特定不能である医薬品も存在した。

「ブランド名」が検出された医薬品において、「規格量」、「濃度」、「剤型」、「屋号」が検出された医薬品の割合は、それぞれ 71.2%、15.7%、94.8%、28.1%であった。基幹語が「ブランド名」である医薬品において、最も多い形態素パターンは「ブランド名＋規格量＋剤型」であった（4882 品目、51.4%）。

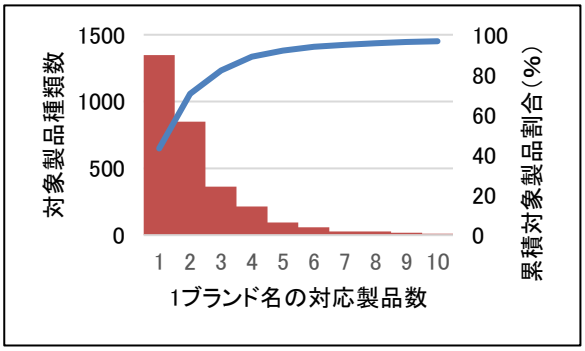

図 1　基幹語がブランド名である医薬品における
1 ブランド名が対応する製品数の分布推移

表 1　基幹語がブランド名である医薬品において
ブランド名のみを指定した場合の代表医薬品コード
（対象製品数の多い上位 10 品目を抜粋）

ブランド語幹	製品数	代表医薬品コード
アムロジピン	165	2171022@XXXX
ロキソプロフェン	91	XX49XXX@XXXX
レボフロキサシン	90	XXXXXXX@XXXX
オルメサルタン	89	2149044FXXXX
ロスバスタチン	85	2189017FXXXX
ピタバスタチン	77	2189016FXXXX
パロキセチン	76	1179041FXXXX
ピオグリタゾン	76	3969007FXXXX
アレルゲン	69	XX9040XA1XXX
クエチアピン	61	1179042@XXXX

X：複数の数字を取りうる
@：複数のアルファベットを取りうる

表 2　基幹語「ロキソプロフェン」における並存形態素を
指定した場合の代表医薬品コード

剤形	規格	製品数	代表医薬品コード
ゲル		3	2649735Q10XX
テープ	（未指定）	48	2649735SXXXX
テープ	100mg	24	2649735S3XXX
テープ	50mg	24	2649735S2XXX
パップ	（未指定）	14	2649735SXXXX
パップ	100mg	13	2649735S1XXX
パップ	200mg	1	2649735S4027
外用ポンプスプレー		2	2649735R10XX
細粒		5	1149019C11XX
錠	60mg	18	1149019F1XXX
内服液	60mg	1	1149019S1038

表 3　基幹語が一般名である医薬品における
並存形態素の組み合わせとその対象製品数

規格量	剤型	濃度	デバイス	屋号	対象製品数
○	○	－	－	○	4058
－	○	－	－	－	1834
－	○	－	－	○	682
－	○	○	－	○	569
－	－	－	－	－	449
－	－	－	－	○	375
○	○	－	－	－	247
○	○	－	○	○	82
－	○	○	－	－	47
○	○	○	○	○	21
○	○	○	－	○	19
○	○	○	－	－	15
○	－	－	－	○	15
－	○	○	○	○	7
○	－	－	○	－	4
－	－	○	－	○	3
－	－	○	－	－	2
－	－	－	○	○	2
○	－	○	○	－	1

さらに形態素を追加指定することにより、代表医薬品コードのワイルドカード範囲は減少し、特定範囲が増大した（表 2）。

一方、「一般名」が検出された医薬品は 8,432 品目（1,149 種類）であり、「一般名」のみで一意に製品を特定できたのは 420 品目（種類数に対して 36.6%）であった。「一般名」が検出された医薬品において、最も多く併存した形態素は「屋号」であった（5,833 品目、69.2%）。基幹語が「一般名」である場合、最も多く存在する形態素パターンは「一般名＋規格量＋剤型＋屋号」であった（表 3）。

4. 考察

我々が構築した医薬品名に特化した形態素解析アルゴリズム DrugMA の判別精度は 99.2%であり、完全なものではない。また、DrugMA は、薬価収載医薬品リストにおける成分名に文字列が完全一致した場合のみ「一般名」と判定する仕様であるため、「アムロジピン錠」（成分名：アムロジピンベシル酸塩）、「ロキソプロフェン Na 錠」（成分名：ロキソプロフェンナトリウム水和物）等は、すべて「ブランド名」として判定される。「ブランド名」を語幹とする医薬品においても、「屋号」の並存率が 28.1%と高率だったのはこのためと思われる。「一般名」の判定基準によって、これらを語幹とする医薬品の割合は大きく変化するものと思われる。

本研究結果より、「ブランド名」や「一般名」によらず、語幹のみの情報でも、一定の割合で一意に医薬品を特定できる可能性が示された。また、並存形態素を追加することにより、医薬品特定率はさらに向上することが確認できた。不具合報告や製品回収等の特殊事例を除き、通常の臨床では、薬価基準収載医薬品コードの先頭 9 桁までの特定が可能であれば、様々な分析に使用可能なレベルでの医薬品特定ができるものと考える。

医療テキスト情報に含まれる医薬品名は、「ブランド名」のみ、あるいは、「ブランド名＋剤型」のみ、「ブランド名＋規格」のみ等、承認医薬品名の一部の形態素が省略されて記載されることが多い。「ブランド名」あるいは「一般名」の語幹ごとに、あらゆる形態素の組み合わせにおける代表医薬品コードに関する辞書を作成することにより、医療テキスト情報に含まれる医薬品名に関する情報から、代表医薬品コードを付与させることが可能になると思われる。