厚生労働行政推進調査事業費補助金(厚生労働科学特別研究事業)
分担研究報告書

自動符号付けアルゴリズムと傷病符号付けチェックの要件
AIを用いた類似度判定についての試行

研究分担者 阿南 誠 川崎医療福祉大学医療福祉マネジメント学部 医療秘書学科 特任教授

研究要旨

類似度判定について、類似度を計算する目的やどのようなデータと比較することによって適した計算方法は変わることになるが、具体的な方法として、標準病名マスターの中から、形態素解析によって単語に分割した後、最も単語数が多いものと類似度を比較することによって、コーディング精度は上げることが可能ではないかと考え試行を行った。対象は、慢性 C 型肝炎、慢性 B 型肝炎、アルコール性肝炎の 3 つの病態から AI ツールの Gemini を用いて判定が可能か否か試行を行い、正しい ICD コードが導かれることを確認した。

A 研究目的

本研究は、自動符号付けアルゴリズムと傷病符号付けチェックの要件を目的として、汎用的な AI ツールを用いて類似度判定が可能であるか試行することが目的である。

B 研究方法

1) 分析対象

慢性 C 型肝炎、慢性 B 型肝炎、アルコール性肝炎の 3 つの病態から検討することとした。

※前提として、慢性肝疾患という共通点、臨床的に共通する特徴があること。疾患の原因に基づく特有の特徴の両方があるという条件で選定した。

2)分析ツール(用いたツール)

研究に用いたツールは、AI ツールのうち、Gemini、形態素解析ツールとしては、MeCab である。

2) 分析方法

具体的な方法として、標準病名マスターの中から、形態素解析によって単語に分割した後、最も単語数が多いもの(ただし、実際には、不必要な単語は排除する必要があると思われるが、今回は配慮しない)、と類似度を比較する。この方法でコーディング精度は上げることが可能か検証する。

C 結果

1) 形態素解析の結果

※Gemini の他、MeCab でも同様の結果を示し、齟齬は発生しなかった。

「慢性 C 型肝炎」 →「慢性」「C 型」「肝炎」

「慢性 B 型肝炎」 →「慢性」「B 型」「肝炎」

「アルコール性肝炎」→「アルコール性」「肝炎」

※C 型は「C」と「型」が分離されることあり。

すなわち、この段階では、形態素解析として、単語の構成として類似しているかどうかは判定出来ても、ICD 分類の条件(定義や意味)は課されていないので、単純に似ているかどうかの判定しか出来ない。そこで、最後は人間か AI 等のチェックが必要となる。

つまり、Mecab を用いた形態素解析から計算出来るのは、見た目の表現「用語・記述の類似度」ということになり、ICD コーディング等の「意味の類似度」はこの段階では判定困難である。例えば、Gemini を用いた場合の解析ステップをみると、形態素解析以降のプロセスを加わることがわかる(表1参照)。

表1. Gemini における解析ステップ

ステップ	詳細
データ収集	各病態に関するテキストデータ（医学論文、教科書の記述、診断基準など）を集める。
形態素解析	集めたテキストを形態素（単語）に分割し、**単語の出現頻度**を数える。
特徴量抽出	**TF-IDF**（Term Frequency-Inverse Document Frequency: 特定の単語の重要度を示す指標）などの手法を用いて、各病態を特徴づける**単語のベクトル**を作成する。
類似度計算	作成されたベクトル間の**コサイン類似度**などの指標を計算し、類似度を数値で示す。

2）Gemini における類似度計算結果について

（1）方法、プロセス

慢性 C 型肝炎、慢性 B 型肝炎、アルコール性肝炎の類似度を、形態素解析で抽出された仮想的な特徴語の出現頻度に基づき、コサイン類似度で計算。

（2）類似度計算の結果

仮想的な医学的記述データに基づく計算結果は以下（3）の通り。

（3）比較ペアコサイン類似度

①慢性 C 型肝炎と慢性 B 型肝炎 {0.853}

②慢性 C 型肝炎とアルコール性肝炎{0.236}

③慢性 B 型肝炎とアルコール性肝炎 {0.211}

すなわち、①の類似度が顕著に高いということがわかる。

（4）計算プロセスについて

各病態に関する医学テキストを形態素解析によって処理し、特徴語とその出現頻度（重み）を抽出。これにより、各病態を多次元空間のベクトルとして表現（詳細な計算方法は割愛）。

表2. 計算プロセス（重み付け等）

形態素 (特徴語),	C型 (VC),	B型 (VB),	アルコール性 (VA)
ウイルス,	5,	7,	0
アルコール,	0,	0,	8
炎症,	3,	3,	4
線維化,	2,	2,	3
肝硬変,	2,	2,	2
治療薬,	4,	3,	0
飲酒,	0,	0,	6
DNA,	0,	3,	0
RNA,	3,	0,	0

3）条件設定を修正し（キーワードの追加、削除）、試みた結果

（1）「ウイルス性」というキーワードを追加した結果

慢性 C 型ウイルス性肝炎と慢性 B 型ウイルス性肝炎の類似度は、以前の計算に基づき、0.853→数値は変わらなかった。

即ち、C 型、B 型でウイルス性ということは認識され影響を与えない。

（2）「慢性」というキーワードを削除した結果

慢性 C 型肝炎と C 型肝炎の類似度は、極めて高い（ほぼ 1.0）という結果であった。このケースもキーワードの慢性（という意味）は認識され影響は与えない。

標準病名マスターにおける ICD コードは、どちらも B18.2 となっており、導き出された結果は正しくコーディングされるといえる。

D 考察

1） 形態素解析による類似度判定

結果の 1）に示すとおり、形態素解析によって、類似度はある一定の比較は可能である。

2） 形態素解析のみでは困難なケース

しかし、病勢、病態、原因等は、ICD コーディングに大きく影響するため、単語の意味を理解しない単純な形態素解析による単語抽出だけでは精度の改善は難しいケースがある。

3） AI ツールの可能性

したがって、標準化が十分ではない、死亡診断書のような記載レベル、入力レベル（必ずしも ICD コーディングのエキスパートの業務ではない）の違いが大きい場合は、その文章の内容（医学的知識や ICD の知識）を理解しなければ類似度を測ることは難しい。しかしながら、ここに AI ツールを導入することによって、精度改善は可能であることが示唆された。

例えば、結果の2）に示すとおり、Gemini による類似度計算結果から推察する限り、十分な満足のいく結果を示した。結果に示すように、慢性 C 型肝炎と慢性 B 型肝炎は非常に高い類似度を示しており。これは、両者が同じウイルス性肝炎であり、「炎症」「線維化」「肝硬変」といった病態の進行に関する共通語を多く含む（と AI が学習している）ためと推察される。一方、ウイルス性の 2 つとアルコール性肝炎の類似度は低い結果となっているが、これは、アルコール性肝炎は「アルコー

ル」「飲酒」といった原因に特有な語の頻度が高く、病因に関する記述が大きく異なり、併せてその語意（疾病への学習）を理解され区別が可能なため、類似度が低く計算されたと推察される。
すなわち、結果の3）で示したとおり、形態素解析を用いるテキストベースの類似度計算において、この二つの用語（ウイルス性、慢性というキーワードの有無）が医学的概念として実質的に同一であると認識されていると推察される。
4）　実装する場合の課題（AI ツールにおける課題）
（1）基本的に類似度を判定するためには、まず、比較する対象（正しい ICD が振られており、必要十分な条件を満たしている「標準的」傷病名であることが必要）を準備、決定する必要がある。
（2）上記の（1）で準備された傷病名とターゲットとした傷病名の類似度を検証することによって、ある程度、同一か否かは検証出来るが、いくつか留意すべき点が考えられる。それは、以下のとおりである。
①利用する AI ツールの違いが出現する可能性がある（Gemini と結果確認のため一部 ChatGPT で解析したが今回の対象について、差異はなかったが、偶然の可能性はある）。
②形態素分析した結果、単語が多くなった場合、類似度が下がる可能性を否定出来ない。すなわち、複雑な傷病名であった場合、専門性等の要因から特徴的な慣習があった場合等、学習が困難になる可能性がある。
③上記の②の影響を考えると、同一の ICD に対して複数の標準的な傷病名（例えば、短い表現と長い表現を準備しておく）を準備する必要があると思われる。改善のためには、汎用的な AI ではなく、医療、医学に適した教師学習を実施された AI が必要になろう。
④データが大量になった場合、解析に時間がかかるであろうことが推察される。

E 結論

今回の慢性肝炎という狭い範囲の試行例ではあったが、AI ツール Gemini で試行した結果を見る限り、現行の AI ツールでも、十分に傷病名における形態素解析結果の単語の医学的意味は理解可能であると思われ、複雑、例外的なケースでなければ、ICD コーディングの一括変換や内容確認には十分に実用レベルにあると考えられた。引き続き、サンプル数を増やした検証も行いたい。

F 健康危険情報

なし

G 研究発表

1.論文発表
なし
2.学会発表
2026 年度、日本医療マネジメント学会学術総会にて発表予定。

H知的財産権の出願・登録状況（予定を含む。）

1.特許取得
なし
2.実用新案登録
なし
3.その他
なし