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

クラスタリングによる不一致パターン分析に基づく既存符号付けアルゴリズムの改善に関する研究

研究分担者 森藤祐史 大阪府立病院機構大阪急性期・総合医療センター
医療情報部 診療情報管理室 総括主査

研究要旨

本研究は、患者調査における ICD 符号付けの精度向上を目的に、辞書判定コードと手付けコードの不一致パターンをクラスタリングにより分析した。分析対象とした不一致 23,498 語(約 39.5%)を K-means 法で 8 クラスタに分類した結果、診療領域ごとに体系的な不一致傾向が確認された。これらの知見に基づき、既存の表記揺れ補正プログラム(xx2.shf)における「Z 付きコードの基本コード置換ルール」の追加や、「目印コード(ZZZZZ 等)」の抽出条件精緻化、さらにはアンマッチ基礎表における参考情報の充実など、ルールベースによる既存プロセスの改善策を提示した。本提案は、高度な自動学習等を用いず、現行の作業アルゴリズムの論理的改善によって符号付け精度の向上と作業負担の削減を目指すものである。

A 研究目的

本研究は、令和 5 年度患者調査までの蓄積である「最新辞書」を用いてもなお、目視による手付け符号付け作業が多数発生している要因を分析することを目的とする。特に、AI 等の新規技術を導入するのではなく、既存の表記揺れ補正 PG や辞書管理、目印コードを用いた個別修正(ステップ 5)等の一連のアルゴリズムにおいて、どのプロセスを論理的に改善すべきかを考察し、実務に即した精度の最適化を図る。また、本研究では、既存の統計処理システムおよび ICD-10(2013 年版)準拠の枠組みの中で、いかに論理的な改善を図るかに焦点を当てている。

B 研究方法

1)分析対象

本研究は、最新辞書によるマッチングの処理を経て参考コードが付与された 59,436 語のうち、目視による手付け符号付け作業により ICD コードが変更された 23,498 語を対象とした。

・対象語数(辞書判定 ICD コードと手付け ICD コードのペア):59,436 件

・両コード不一致語数:23,498 件(約 39.5%)

クラスタリング前の対象語数の分布は図 1 のとおり。また、変更理由のコメント欄に符号付け作業時の記載がある 784 語から、上位グループは下表(表 1)のとおり。

表 1 不一致理由の上位グループ

確認済みの簡易コメント(235 語)
年齢条件(16 歳未満など)(58 語)
初筆(初回記載)に関する判断(24 語)
「?」のある確認コメント(20 語)
外傷性切断の術後か切断術後かの判断(18 語)
外傷関連(18 語)
年齢によるコード分岐(例:16 歳未満=L20.8、16 歳以上=L30.9)(15 語)
部位の明記不足で局所性と判断(13 語)
股関節の詳細不明疾患(9 語)

辞書コードは一般的な ICD コードを付与するが、手付けでは下表(表 2)のように、より詳細なコー

ドを選択されたケースが多い。

表２

傷病名	辞書 ICD	手付け ICD
サルモネラ症	A02.9	A02.0
ネズミチフス菌腸炎	A09.9	A02.0
腸管出血性大腸感染症	A09.0	A04.3
急性細菌性腸炎	A09.9	A04.4

そのほか、「NOS コード（詳細不明）への変更（部位や病態が不明確な場合に手付けで NOS を選択）」、「年齢や外傷の有無など、辞書では拾えない条件分岐が必要なケース」、「術後状態や後遺症など、文脈依存の判断が必要なケース」などのパターンがある。

これにより、ICD コードの不一致パターンは、類似性に基づくグルーピングが有効な可能性がある。

２）分析手法

辞書コードと手付けコードのペアを数値化し、K-means 法によるクラスタリングにより差異特徴として抽出した。患者調査の統計公表における最も詳細な単位は「傷病基本分類（ICD-10 の 3 桁項目および 4 桁細分類項目）」として設定されており、この分類体系に 5 桁目の情報は含まれない。しかし、符号付けの実務プロセスにおいては、部位詳細コードや条件判定のための目印コード等の 5 桁目の情報が実務上の解像度として介在している。本研究は自動符号付けアルゴリズムの最適化を主眼としており、4 桁以下に丸めた場合には、Cluster 0（Z コードの基本コードへの置換）や Cluster 1（部位詳細コードの付与）に代表される「実務上で発生している不一致パターン」が構造的に観察不能となる。そのため、本分析では実務上のボトルネックを精緻に特定する目的から、5 桁解像度での解析を採用した。

ICD コードは文字列のデータであるが、厳密なグループと階層の構造を持ち、コードの並び自体が医学的近接性を反映しているため、医学的類似度をユークリッド距離として近似的に抽出できる発見的手法として妥当であると判断した。各ペアを 2 次元の数値ベクトル（辞書コード、手付けコード）に変換したラベルエンコーディングにより数値化する。クラスタ数はエルボー法によって選定する。不一致の原因については、既存の「表記揺れ補正処理（xx2.shf）」や「患者独自ルール（ステップ 1～3）」との整合性の観点から整理した。分析にあたっては、ICD コード間の差異を客観的に評価するため、一時的に数値化した。これは実務上の既知の不一致傾向を論理的に整理する目的であり、自動学習モデルの構築を目的としていない。

C 研究結果

最新辞書によるマッチングを経て不一致となった語に対し、K-means 法を用いたクラスタリングを実施した結果、以下の通り診療領域ごとの体系的な特徴が抽出された。

１）クラスタリングの実施と分類状況

分析対象とした不一致語のうち、実運用における出現頻度を考慮するため、語彙単位のデータをレコード単位に再構成し、欠損値の除外や正規化処理を行なった。30,020 レコードが対象となった。シルエット係数（SC : Silhouette coefficient）はクラスタ数 6 で最大値を示したが、クラスタ数 8 においても SC =0.52 と高い値を維持していた。クラスタ数の決定にあたっては、クラスタ内分散（WCSS：Within-Cluster Sum of Squares）の減少傾向に加え、実務上の診療ドメイン（循環器、外傷等）の分離状況が最も整合的であったことから、クラスタ数 8 を採用した。（図２、表３…図表中の WCSS は 10^6単位で表示した）

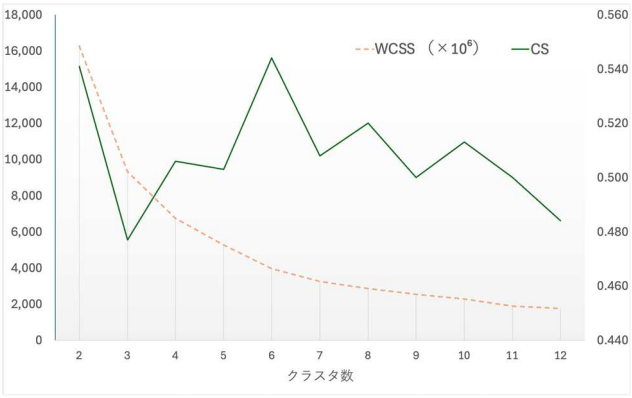


図２　エルボー法およびシルエット係数

表３　エルボー法およびシルエット係数

クラスタ数	WCSS	SC
2	16,265	0.541
3	9,313	0.477
4	6,742	0.506
5	5,259	0.503
6	3,960	0.544
7	3,251	0.508
8	2,856	0.52
9	2,530	0.5
10	2,269	0.513
11	1,885	0.5
12	1,758	0.484

各クラスタの名称および分類語数は下表（表４）。クラスタ名称は、当該クラスタにおいて最も頻出した不一致パターンに基づいて付与しており、必ずしも単一診療領域のコードのみで構成されるものではない。

表４　クラスタ名称および分類語数

Cluster 0：循環器疾患	5,363 語
Cluster 1：外傷	5,841 語
Cluster 2：感染症	3,864 語
Cluster 3：泌尿器・産科関連	3,321 語
Cluster 4：眼科・精神疾患	2,807 語
Cluster 5：皮膚・筋骨格疾患	5,322 語
Cluster 6：妊娠・管理関連	1,800 語
Cluster 7：その他	1,702 語

２）クラスタ別の代表的な不一致パターン

各クラスタにおいて、形態素分析による判定（R5_AC）と手付けコードの間で発生した代表的な不一致ペアとその差異の特徴は以下の通りである（表５より抜粋）

- Cluster 0（循環器疾患）：
  事例：I50.9Z（形態素分析）→ I50.9（手付け）
  特徴：循環器や呼吸器疾患における**「Z 付きコード」の基本コードへの置換**が目立つ。
- Cluster 1（外傷）：
  事例：S06.5B → S06.5、S32.0F → S32.0
  特徴：骨折部位の詳細な分類（部位コードの付与）による不一致が主である。
- Cluster 2（感染症）：
  事例：B02.9 → B02.2、D27.Z → D27
  特徴：原因菌やウイルス種の詳細差、および腫瘍コードの詳細化による差異が確認された。
- Cluster 3（泌尿器・産科関連）：
  事例：U99 → O34.2（妊娠管理関連）、M86.9B → M86.97（筋骨格）
  特徴：妊娠合併症や管理に関する文脈依存の判断、および筋骨格系の詳細部位差が見られる。
- Cluster 4（眼科・精神疾患）：
  事例：H40.9 → H40.1、F03 → G30.9
  特徴：白内障や緑内障、認知症関連における詳細分類の選択に差異が生じている。
- Cluster 5（皮膚・筋骨格疾患）：
  事例：L30.9 → L85.3、M48.0 → M48.0
  特徴：皮膚疾患の分類体系の差や、脊椎疾患の詳細部位の選択による不一致が多い。
- Cluster 6（妊娠・管理関連）：
  事例：R54 → H35.3、R58.Z → I61.0
  特徴：R コード（症状・徴候）から特定の疾患コード（眼疾患や循環器疾患等）への置換が多い。
- Cluster 7（その他）：

事例：J12.8F → U07.1（COVID-19 関連）
特徴：呼吸器疾患の詳細差や、特定の流行性疾患への修正が顕著である。

3）不一致要因の定性的分析

変更理由のコメントが付与されていた 784 語の分析からは、次の要因が上位に挙げられた。「年齢条件（16 歳未満等）によるコード分岐」、「初筆（最初に記載された傷病名）の優先判断」、「部位の明記不足による局所性判断の要否」、「？を含む不明瞭な記載に対する目視確認」。これらの結果は、単なる辞書の網羅性不足だけではなく、既存の「表記揺れ補正」や「目印コード」のロジック、および作業者の判断基準を精緻化する必要があることを示唆している。

D 考察

本研究のクラスタリング結果から、特定の診療領域において体系的な不一致傾向が確認された。形態素解析において、特定の診療領域（外傷の部位や感染症の菌名など）が未知語（辞書未登録語）として適切に処理されず、詳細不明コード（NOS）に流れてしまうという既存の形態素解析処理の限界が、クラスタ分析によって可視化された。これらの知見に基づき、既存の符号付けアルゴリズムおよび作業プロセスの改善に向けた考察を行う。

1）クラスタリングに基づく診療領域別の不一致要因分析

各クラスタの不一致パターンには、以下の通り診療領域固有の特徴が見られた。

- Cluster 0（循環器疾患）： 末尾に「Z」が付与されたコードが基本コードに置換されるケースが多く、詳細度よりも基本分類を優先する実務上の判断が反映されている。
- Cluster 1（外傷）： 骨折部位の詳細差異が主因である。これは傷病名に部位情報が不足している場合、既存のルールに従い詳細不明または代表的な部位を選択する手作業が発生しているためである。
- Cluster 2（感染症）： 原因菌やウイルス種の違いによるコード差異が頻出している。これらは辞書側の網羅性や、臨床病名と統計分類上の名称の乖離に起因する。
- Cluster 3（泌尿器・産科関連）： 妊娠合併症等の詳細分類において、患者の背景情報（妊娠週数等）を考慮した手付け判断が優先されている。
- Cluster 4（眼科・精神疾患）： 白内障の分類（老年性等）や認知症関連において、年齢等の条件分岐を伴うコード選択に差異が生じている。
- Cluster 5（皮膚・筋骨格疾患）： 皮膚疾患の分類体系や脊椎疾患の詳細部位において、辞書コードよりも手付けによる詳細化が図られている。
- Cluster 6（妊娠・管理関連）： 妊娠管理や産科合併症において、調査票の「受療の状況」等の文脈を反映したコード修正が行われている。
- Cluster 7（その他）： 複合診断や後遺症に関し、初筆優先ルール（原則として一番左側に記載の傷病名を選択する）に基づく修正が多く見られる。

本分析は ICD の章構造を一定程度再現しているが、これは診療領域ごとの対策を可能にする意図的な設計である。一方で、Cluster 7 における J コード（呼吸器）から U コード（特殊目的）への移行や、Cluster 3 における他領域から O コード（産科）への修正などは、文字列の近さを超えた「臨床的優先順位」や「実務上の判断ルール（初筆優先）」という章横断的な不一致構造を可視化している。

2）既存アルゴリズムおよびプログラムへの改善提案

分析で得られた頻出パターンを、既存の「職員独自作成プログラム（xx2.shf）」と「AC システム」の役割分担に基づき反映させることで、精度の向上と作業の効率化を図る。

- 表記揺れ補正 PG（xx2.shf）による前処理の拡張：Cluster 0 や 5 等で確認された「Z 付きコード」や「代表的な表記揺れパターン」を、既存プログラムの「漢字・記号変更ルール」や「誤字修正一覧」シートに直接定義することを提案する。特に、意味が通じる範囲での最小文字数による置換設定を精緻化することで、後工程の形態素解析に回る前の「完全一致マッチング率」を論理的に向上させることが可能である。
- 辞書および AC システムにおける参考情報付与の改善：xx2.shf を通過したデータに対し形態素解析を行う「AC システム」において、Cluster 2（感染症）等の原因菌による差異に対応するため、既存の辞書へのキーワード追加や優先順位の最適化を行う。また、既存の「アンマッチ基礎表（Excel）」作成時に、特定のキーワードが含まれる場合に「過去の修正事例」を自動提示する仕組みをマクロ等の処理工程に追加・改善する。
- 目印コード（ZZZZZ 等）の抽出条件の精緻化：Cluster 1（外傷）の部位不明や、Cluster 3（産科関連）の母子判定など、傷病名のみではコード決定が困難なケースに対し、特定のキーワード（「右上？骨」等）に基づき「目印コード（ZZZZZ、O コード等）」を自動付与する論理を強化する。これにより、ステップ 5（個別修正）において確認が必要な対象をより正確に抽出し、不適切なコードがそのまま計上されるリスクを低減できる。

3）手付け符号付け作業の標準化と教育への応用

AI 等による自動化に頼らずとも、不一致が発生しやすい領域を明確化することで、作業者の判断アルゴリズムを平準化することが可能と考えられる。

- 標準作業手順（SOP）の再定義：Cluster 0, 1, 5 等で抽出された詳細分類の不一致例を、実務担当者向けの「符号付け手順書」や「引き方のコツ」の具体的例示としてフィードバックする。
- 判断ルールの徹底： ステップ 4 における「詳細不明コード（NOS）の回避」や「初筆優先」といった既存の原則を、分析で判明したエラーリスクの高い具体例を用いて再徹底することで、人的なコーディング精度の向上を図る。

E 結論

本研究では、形態素解析による判定 ICD コードと手付け ICD コードの不一致パターンをクラスタリングにより分析し、診療領域ごとの特徴を明らかにした。その結果、循環器疾患、外傷、感染症など特定領域で体系的な不一致傾向が確認された。これらの知見は、既存の「表記揺れ補正 PG」の定義更新、目印コードの抽出ロジックの改善、および作業手順の標準化といったルールベースの対策に直接活用可能である。現行の手法を維持しつつも、一連の作業アルゴリズムを論理的に洗練させることで、符号付け精度の向上と実務の効率化が十分に可能であることが示唆された。

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

なし

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

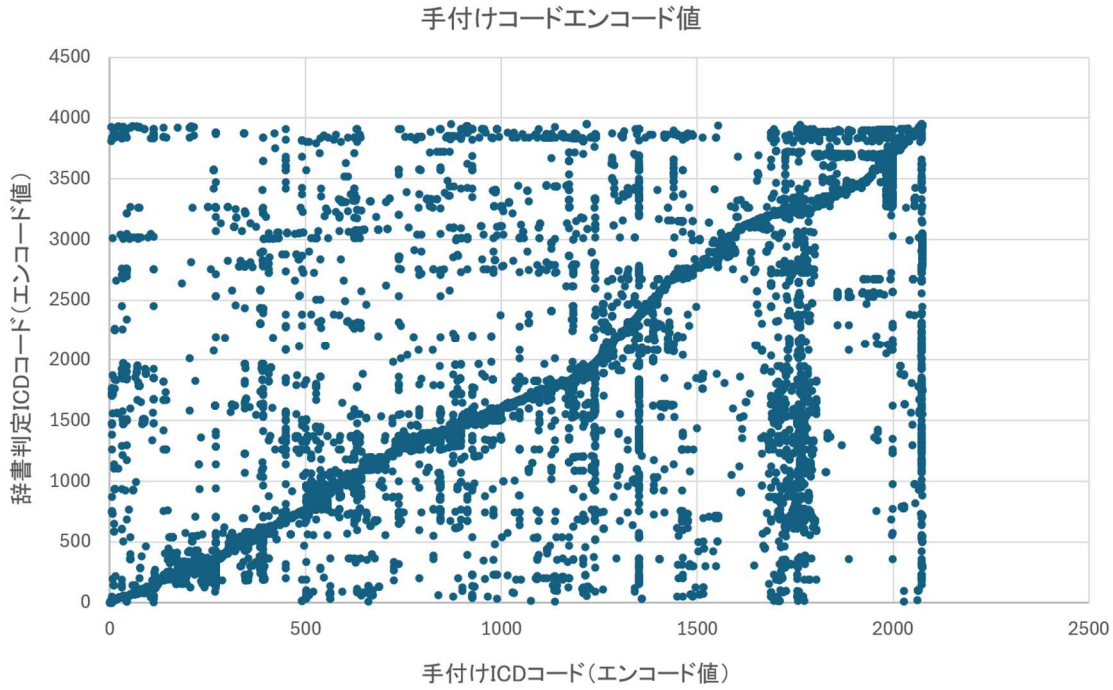


図１　クラスタリング前の対象語数

（形態素解析による判定 ICD コードと手付け ICD コードのペア）の分布

表５　クラスタ別の代表的な上位パターン（上位５例）

クラスタ No	クラスタ内の語数	代表ペア		語数	差異の特徴
		R5_AC	手付け		
Cluster0：循環器疾患	5363	I50.9Z	I50.9	222	循環器疾患の詳細差
		I63.9	I63.5	133	循環器疾患の詳細差
		J69.0Z	J69.0	133	呼吸器疾患の詳細差
		K63.5B	K63.5	85	その他
		I63.9	I63.3	79	循環器疾患の詳細差
Cluster1：外傷	5841	S06.5B	S06.50	192	外傷・損傷の部位差
		S32.0F	S32.00	151	外傷・損傷の部位差
		S72.0C	S72.00	100	外傷・損傷の部位差
		S72.1C	S72.10	78	外傷・損傷の部位差
		U97	R69	76	妊娠・管理関連
Cluster2：感染症	3864	B02.9	B02.2	85	感染症の原因差
		D27.Z	D27	70	腫瘍コードの詳細差
		D48.7	D48.1	53	腫瘍コードの詳細差
		D64.9Z	D64.9	48	腫瘍コードの詳細差
		D32.9	D32.0	44	腫瘍コードの詳細差
Cluster3：泌尿器・産科関連	3321	N18.9	N18.5	60	その他
		M86.9B	M86.97	58	筋骨格疾患の詳細差
		U99	O34.2	55	妊娠・管理関連
		U99	O42.9	39	妊娠・管理関連
		U99	O36.4	34	妊娠・管理関連
Cluster4：眼科・精神疾患	2807	H40.9	H40.1	57	その他
		H02.0Z	H02.0	55	その他
		H26.9	H25.0	41	その他
		H40.9	H40.5	41	その他
		F03	G30.9	39	その他
Cluster5：皮膚疾患・筋骨格疾患	5322	L30.9	L85.3	164	皮膚疾患の分類差
		M48.0	M48.06	130	筋骨格疾患の詳細差
		M17.9	M17.1	98	筋骨格疾患の詳細差
		M17.9	M17.0	83	筋骨格疾患の詳細差
		M81.9	M80.98	82	筋骨格疾患の詳細差
Cluster6：妊娠・管理関連	1800	R54	H35.3	79	その他
		R54	H25.0	35	その他
		R58.Z	I61.0	35	その他
		R56.8	H81.1	34	その他
		R54	H25.9	30	その他
	1702	J12.8F	U07.1	378	呼吸器疾患の詳細差

Cluster 7：その他		M62.5	U07.1	79	筋骨格疾患の詳細差
		J12.8H	U07.1	53	呼吸器疾患の詳細差
		J12.8H	U09.9	51	呼吸器疾患の詳細差
		J18.9	U07.1	33	呼吸器疾患の詳細差