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

標準病名マスターを照合源とする符号照合アルゴリズムを用いた患者調査の傷病名符号化の最適化と効率化に資する並行照合手段の試作と評価

研究分担者　荒井　康夫　北里大学 未来工学部データサイエンス学科　准教授
研究分担者　西本　　寛　JA長野厚生連 佐久総合病院 総合医療情報センター長

研究要旨

患者調査は、統計法に基づく基幹統計である患者統計を作成し、医療施設を利用する患者の傷病状況等を明らかにするための調査である。患者調査では、医療機関が調査票の傷病名欄に記入した傷病名に ICD-10（疾病及び関連保健問題の国際統計分類第 10 版）コードが付与される。傷病名は自由記述形式で記入されるため、表記揺れや構造的多様性を含む。このため、辞書照合で処理されず用手的確認を要する語が発生する。本研究は、患者調査の傷病名符号化プロセスを前提として、用手的確認を支援する補助的検証手段の試作・評価を目的とした。一般財団法人医療情報システム開発センター（MEDIS-DC）が公開する ICD10 対応標準病名マスターを照合源とし、符号照合アルゴリズム PCMA（Patient-survey Coding Match Algorithm；患者調査符号照合アルゴリズム）および照合指数 PCMI（Patient-survey Coding Match Index；患者調査符号照合指数）を試作した。令和 5 年患者調査で確定した ICD-10 準拠統計基準コードを参照基準とし、3 桁分類単位で一致した場合を本研究における正答と定義した。主指標対象 111,984 語における PCMA 第 1 位候補の確定符号との一致率は 85.09％であった。本研究の正答率は確定符号との一致率であり、独立した医学的正解との一致率ではない。形態素解析と比較可能であった 59,375 語では、PCMA は形態素解析とは異なる確認手がかりを提供し得ることが示された。辞書照合で処理されなかった語 75,101 語のうち、3 層化可能であった 75,068 語を形態素解析と PCMA の出力関係により整理し、用手的確認の対象重点化に資する構造を示した。PCMA は確定符号を自動決定するものではなく、候補提示、判定根拠の明示、判断負荷の軽減可能性に資する補助的検証手段として位置づけられる。独立参照基準の構築および過年度データによる外的妥当性検証は今後の課題である。

A. 研究目的

患者調査は、統計法に基づく基幹統計であり、医療機関から無作為抽出された施設で調査日に受療した患者の傷病情報を、世界保健機関（WHO）が定める『疾病及び関連保健問題の国際統計分類（ICD）第 10 版』に基づき集計する[1]。傷病名から ICD-10 コードへの変換（以下「傷病名符号化」）の正確性と再現性は、本基幹統計の品質を担保する。

患者調査において、医療機関が調査票（紙または電子）の傷病名欄に記入する傷病名は、自由記述形式である。患者調査では、紙調査票では傷病名欄に傷病名が自由記述され、電子調査票では電子カルテ情報等から得られた情報を登録する運用も想定さ

れている。記入される傷病名は、診療上の傷病概念を表現する自由文であるため、表記揺れ（異字・誤字・略語・全角半角混在）、構造的多様性（非傷病名・並列記載・因果関係記載）、否定・両側性・多発性等の医学的属性を含む。本研究において「非傷病名」は、修飾語のみ、部位のみ、診療科名、後置語、なし表現などで構成され、傷病名として成立しない語を指す。このため傷病名辞書との照合のみでは大部分の語を符号化できず、形態素解析または手付けによる符号化を要する。患者調査の傷病名符号化プロセスは、辞書符号付け、形態素解析が候補を提示した語に対する手付け、形態素解析により候補が提示されなかった語に対する用手的な手付け符号化の 3 工程で構成される。

手付け符号化は、ICD-10 索引表および疾病、傷害及び死因の統計分類提要 ICD-10 準拠第 1 巻（内容例示表）を参照しながら、自ら ICD-10 コードを探索する学術的かつ調査的な作業である。手付け符号化の対象範囲では、形態素解析の出力（候補提示の有無、提示された候補の妥当性）が当該作業の手がかりの有無を規定する。形態素解析が候補を提示しなかった語に対する手付けは、手がかりのない状態で行われる。

本研究の目的は、患者調査の傷病名符号化プロセスに対する並行照合の手段として、標準病名マスターを照合源とする符号照合アルゴリズム PCMA（Patient-survey Coding Match Algorithm）および照合指数 PCMI（Patient-survey Coding Match Index）を試作することである。試作後、令和 5 年患者調査の出現した傷病名を試料として、当該プロセスの最適化および効率化への資する程度を評価する。本研究において「最適化」は、患者調査における傷病名符号化プロセスを前提として、標準病名マスターとの照合結果に基づき、適合度の高い符号候補を選択するための判断材料を提示し、用手的な確認を支援することで、符号化の正確性、判定根拠の明示性および再現性の向上に資することを指す。一方、「効率化」は、目視作業を要する語の量的縮減可能性および判断負荷の軽減を指し、所要時間の短縮は本研究の検討対象に含めない。本研究は、患者調査プロセスを変更・代替するものではない。確定符号とは独立した照合ロジックを並行して提供することにより、第二の判定根拠を明示性・再現性のある形で付与する。これにより、手付け符号化が必要な語に対する調査・探索作業の軽減可能性を提供する。

B. 研究方法

1）データソースと評価分母

対象データは、令和 5 年患者調査の傷病名である[2]。語数の単位は得られた種類数（type 数。重複を除いた傷病名の種類数）であり、延べ出現数とは区別する。原データ 112,133 語から評価上の除外語を控除し、主指標対象 111,984 語、参考指標対象 110,670 語の 2 つの評価分母を定義した（除外項目の詳細は表 2）。PCMA の第 1 位候補とは、PCMI が最上位の候補を指す。

2）PCMA の構成と照合源

PCMA は、一般財団法人医療情報システム開発センター（MEDIS-DC）公開の ICD10 対応標準病名マスター（Ver.5.17[3]、以下「標準病名マスター」）を照合源とする Python 実装のルールベースアルゴリズムである。標準病名マスターは臨床現場で用いられる病名集であり、傷病名の臨床記載と ICD-10 を架橋する照合源として整合する。

標準病名マスターを照合源として採用した理由は、次の 3 点である。第一に、本マスターは厚生労働省標準規格 HS005 に位置づけられた公的標準であり、医療現場で広範に利用されている。第二に、各標準病名に ICD-10 準拠統計基準のコード（ICD-10 コード）が直接付与されている。本マスターを照合源

とすることにより、文字列照合の結果から ICD-10 コードの候補を直ちに提示できる構造とした。第三に、本マスターは臨床で用いられる病名表現を網羅している。これにより、本マスターを照合源とする照合の到達範囲を確保した。なお、本マスターは、基本病名に修飾語を付加することにより、多様な病態表現に対応する構造を有する。

本研究において、生成 AI および機械学習を主要な照合手段として採用せず、ルールベース照合を選択した理由は、次の 3 点に集約される。第一は再現性である。基幹統計は経年比較を前提とする。このため、同一入力に対して同一出力が得られる必要がある。第二は説明可能性である。基幹統計の符号化では、判定根拠が第三者により検証可能でなければならない。第三は判定根拠の明示性である。本研究のルールベース照合では、照合ルールと点検ルールが判定根拠の明文化に直接対応している。これらのルールは、規程化・公開・改訂が可能である。

PCMA の処理は、(1) 正規化（Unicode 正規化 NFKC、全角・半角統一、異字誤字変換、修飾語の識別と除去）、(2) 標準病名マスターの索引用語との照合および PCMI 算出、(3) 上位候補の順位付け、(4) 概念点検（ICD-10 内容例示表に基づく医学概念と章・サブ分類の整合性判定）、の 4 段階で構成される。複合病名は句読点等を区切りとして並列分割し、先頭パートを優先する。ただし先頭パートが第ⅩⅧ章（症状・徴候）または第ⅩⅪ章（保健サービス）であり、かつ後続に第Ⅱ章（悪性新生物）等の確定診断名が含まれる場合は、後者を優先する。各傷病名に対し、PCMI が高い順に上位 5 語を候補として出力する。本研究では、PCMA が上位 5 語までを提示する設計とした。これにより、最上位候補のみを返すのではなく、後段の手付け符号化における判定の選択肢を確保した。本研究の評価では、第 1 位候補に紐づく ICD-10 コードを PCMA の判定結果とした（表 3）。PCMA は原則として全評価対象に対して候補を提示できる設計であるが、(a) データ例外（記号・数字・空文字のみで構成される語、14 語）、(b) 非傷病名（135 語）、(c) 標準病名マスターに収載されるが当該病名単独では ICD-10 コードが一意に定まらない語または PCMA の照合で有効な候補が選定されなかった語（ICD 未設定、116 語）の 3 区分に該当する語については、構造上候補を提示しない。これら計 265 語のうち、データ例外および非傷病名は主指標対象から既に除外され、ICD 未設定は参考指標対象から追加除外される（表 2）。

標準病名マスターを照合源として用いる際の注意点は、次の 2 点である。第一に、修飾語が付加された病名表現に対して、基本病名に付与された ICD-10 コードが必ずしも妥当ではない場合がある。第二に、文字列近似照合は文字一致度を最大化する候補を選ぶ。これは、臨床概念上の整合性を保証するものではない。文字列の一部相違や語順差により、文字列上は近いが臨床的には別概念である候補が選ばれる余地が残る。本研究では、これらの注意点に対応するため、照合（PCMI による候補順位付け）と概念点検（ICD-10 内容例示表に基づく整合性確認）を別工程として配置した。両工程の分離により、文字列上は高一致であっても臨床概念上に矛盾を含む候補を識別する経路を確保した。

複数部位の記載を含む傷病名に対しては、単一部位の標準病名に紐づく標準病名マスター側ICD-10 コードでは妥当な符号を選択できない場合がある。本研究では、これらの語に対して多部位対応の符号への置換を試行的に導入した。

3）PCMI と重み係数の最適化

PCMI は、出現した傷病名と候補病名の類似度を 0〜1 の範囲で表す指標である。共通文字率、非共通ペナルティ、語順整合度、索引用語長係数、2-gram 一致率、形態素類似度の 6 観点の重み付き線形結合に、ICD-10 章・サブ分類の整合性を表す

互換性係数 f_compat を乗じ、理論最大値で正規化する。

$$PCMI = clamp(\Sigma w_i s_i / S_max \times f_compat^{w_7}, 0, 1)$$

重み係数 w_1〜w_8は、Optuna の CmaEsSampler を用いた共分散行列適応進化戦略（CMA-ES）[5][6]により、第 1 位候補の正答率を最大化する組合せを探索した（乱数シード固定）。PCMI の構成要素と算出例を表 4 に示す。

4）評価指標

参照基準は、患者調査の傷病名符号化プロセスで付与された確定 ICD-10 準拠統計基準コード（以下「確定符号」）である。正答は、PCMA の第 1 位候補の ICD-10 コードが確定符号と ICD-10 における 3 桁分類以上で一致した場合と定義した。ICD-10 における 3 桁分類は、厚生労働省「疾病、傷害及び死因の統計分類」の基本分類表と整合する公的統計の主要な表章単位である。本研究の正答率は確定符号との一致率であり、独立した医学的正解との一致を評価するものではない。

形態素解析が候補を提示した語に対する手付けの対象における PCMA と形態素解析との比較には、McNemar 検定（Yates 連続補正、α=0.05）を用いた。95%信頼区間は Wilson 法により算出した。実装は Python 3.13 を用い、主要ライブラリは pandas、numpy、SudachiPy、Optuna である。

本研究において「一致」と「正答」を次のように区別する。「一致」は、2 つの出力（PCMA の出力と形態素解析の出力、または PCMA の出力と患者調査の確定符号）が同一の ICD-10 コードを指す状態を表す。「正答」は、PCMA の第 1 位候補の ICD-10 コードが患者調査の確定符号と ICD-10 における 3 桁分類以上で一致した場合を指す。

5）評価の二段構成と当初設計

本研究は当初、手付け符号化を要する語（辞書符号付けで処理されなかった語、75,101 語）に対する PCMA の正答率評価を想定して設計した。そのなかで、辞書符号付け（患者調査の過去実績で正答として確定した範囲）に対する評価は、PCMA の性能を確認する明確な参照点として併せて試行した。研究実施中、PCMA の選択した索引用語と出現した傷病名が完全一致しながら、確定符号と一致しない事例を確認した。これにより、標準病名マスターと患者調査の符号付け基準の間に運用方針の差異（以下「符号化方針差」）が存在することが判明した。これは PCMA の文字列照合上の誤りではなく、両体系の符号付け基準の差異に起因する。

この差異への対処として、本研究では評価分母を二段構成とした。主指標対象（111,984 語）は出現した傷病名から評価上の除外語のみを控除したもの、参考指標対象（110,670 語）はそこから符号化方針差および ICD 未設定・参照基準不明を追加除外したものである（表 2）。符号化方針差の事例は PCMA の誤りでなく両体系の運用方針差として扱う。

（倫理面への配慮）

本研究は患者調査の出現した傷病名のみを対象とし、個人を直接識別できる情報を含まない集計済み統計データの二次的解析である。本研究は「人を対象とする生命科学・医学系研究に関する倫理指針」の研究定義に該当しないと判断し、所属機関倫理審査委員会への申請は行わなかった。標準病名マスターは、MEDIS-DC への利用申請および利用許諾の取得のうえで使用した。

C. 研究結果

1）主指標と工程別正答率

主指標対象（111,984 語）における PCMA 第 1 位候補の正答率は 85.09%（95,288／111,984、95%

CI：84.88–85.30）であった。工程別の正答率は、辞書符号付けで 92.03%、形態素解析が候補を提示した語に対する手付けで 83.48%、手付け符号化で 74.87%であった（表 1、表 5）。

2）形態素解析が候補を提示した語に対する手付けの対象における PCMA と形態素解析の比較

形態素解析が候補を提示した語に対する手付けの対象 59,408 語のうち、形態素解析側で第 1 位候補が出力された 59,375 語（差 33 語は比較不能のため除外）について、両者の出力 ICD-10 コードと確定符号との関係を集計した（表 6）。両者の出力が ICD-10 における 3 桁分類以上で一致した語は 41,754 語（70.32%）であった。確定符号との正答率は、PCMA が 83.49%、形態素解析が 72.02%であり、PCMA が 11.47 ポイント上回った（McNemar 検定 χ^2=3,165.39、p<0.0001）。

3）手付け符号化における判断負荷を軽減する可能性がある語の 3 層構造

辞書照合で処理されなかった語（形態素解析が候補を提示した語に対する手付けの対象 59,408 語と手付け符号化の対象 15,693 語の合計 75,101 語、主指標対象に対し 67.06%）を、「手付け符号化における判断負荷を軽減する可能性がある語」と定義した。当該語における形態素解析と PCMA の出力関係に基づき、3 層に層別化した（表 7）。

第 1 層は、形態素解析と PCMA が一致した符号を提示した語（41,754 語、当該語のうち 55.60%）である。形態素解析（言語処理アプローチ）と PCMA（標準病名マスター照合アプローチ）という異なるアプローチが同一の ICD-10 コードに到達した語であり、当該層の正答率は 93.03%であった。

第 2 層は、形態素解析と PCMA がいずれも符号を提示するが両者が一致しない語（17,621 語、当該語のうち 23.46%）である。両者を手がかりとして手付け符号化に資する語であり、PCMA 側は PCMI が候補病名の文字列上の近似度を表す。

第 3 層は、形態素解析が符号を提示せず、PCMA のみが符号を提示する語である。形態素解析が手がかりを与えなかった語にも PCMA が候補を提示できることにより、手がかりがなかった語における手付け符号化を軽減する可能性が生じる。PCMA 側は PCMI が候補病名の文字列上の近似度を表す。

4）当該語における不正答内訳

第 1 層 41,754 語における不正答は 2,910 語（6.97%）であった。当該不正答の内訳は、アルゴリズム改善余地（残余として算出）が 2,171 語（74.60%）、符号化方針差に起因する語が 696 語（23.92%）、ICD 未設定の語が 28 語（0.96%）、参照基準不明の語が 15 語（0.52%）であり、残存リスクの大部分は PCMA のアルゴリズム改善余地に帰属する（表 10）。

5）PCMI 閾値別の正答率

PCMI の閾値別の該当語数と正答率を集計した（図 1）。閾値が高いほど該当語数（カバー率）は単調に減少する一方で、正答率は閾値 0.50〜1.00 の範囲で 88.6〜89.2%の狭い帯域に留まり、明確な単調関係は確認されなかった。この結果は、PCMI が候補順位付けの指標として最適化されていることを反映する。したがって、PCMI 値そのものを点検対象選別の主指標として用いることは適切でない。手付け符号化における判断負荷を軽減する可能性がある語の層別化は、PCMI 値ではなく「形態素解析と PCMA の出力関係」という離散的根拠を主とする。

6）標準病名マスターと患者調査の符号付け

PCMA が選択した索引用語と出現した傷病名が完全一致しながら、確定符号と不一致であった語は 1,184 語（符号化方針差と総称）であった。内

訳は、索引一致・参照基準相違 746 語、左右差不問・参照基準相違 438 語である。これらは文字列照合上の誤りではなく、標準病名マスターと患者調査の符号付け基準の差異に起因する。これらの語は参考指標対象から除外した。

D. 考察

本研究の主要所見は次の 3 点である。第一に、PCMA が患者調査の確定符号に対し、主指標対象における第 1 位候補正答率 85.09%を示したこと。第二に、形態素解析が候補を提示した語に対する手付けの対象における直接比較で、PCMA の正答率が形態素解析を 11.47 ポイント上回ったこと。第三に、手付け符号化における判断負荷を軽減する可能性がある語（75,101 語）を 3 層に層別化し、形態素解析と PCMA が一致した符号を提示する第 1 層（41,754 語、当該層正答率 93.03%）を識別したことである。

（最適化への資する程度）

PCMA は、確定符号とは独立した照合ロジックに基づき、全評価対象に対して判定結果を提示できることが確認された。判定根拠は文字列照合（PCMI による候補順位付け）と概念点検（ICD-10 内容例示表に基づく医学概念整合性確認）から構成され、明示性・再現性のある形で第二の判定根拠を提供する。辞書符号付けで確定した語に対する PCMA の高い正答率は、PCMA が辞書による自動付与とほぼ整合する判定を再現し得ることを示す。一方で、標準病名マスターと患者調査の符号付け基準の差異（符号化方針差）が研究実施中に判明し、PCMA の文字列照合上の正しさが両体系の運用方針差により評価される構造的制約が確認された。

（効率化への資する程度）

手付け符号化は、ICD-10 索引表および ICD-10 内容例示表を参照しながら、自ら ICD-10 コードを探索する学術的かつ調査的な作業である。PCMA の候補提示は、当該作業に対する手がかりを与える。第 1 層（形態素解析と PCMA が一致した符号を提示する語）では、異なるアプローチが同一の符号に到達したことが手付け符号化を軽減する可能性の根拠となる。第 2 層では両者の候補が手がかりとして併存し、PCMA 側は PCMI が候補病名の文字列上の近似度を表す。第 3 層は、形態素解析が候補を提示しなかった語、すなわち手がかりがなかった語である。PCMA は当該語にも候補を提示でき、PCMI が候補病名の文字列上の近似度を表すことにより、手がかりがなかった語における手付け符号化を軽減する可能性が生じる。本研究は作業時間を測定していないため、効率化の効果は実測値ではなく、目視作業を要する語の量的縮減可能性および判断負荷の軽減として解釈する。

（辞書符号付けに対する本研究の立場）

辞書符号付けは患者調査の過去実績において正答として確定した範囲であり、本研究はこれに対して「別のアプローチ」を要するものとして提示するものではない。本研究において「異なるアプローチ」とは、形態素解析（言語処理アプローチ）と PCMA（標準病名マスター照合アプローチ）の対比のみを指す。当初は手付け符号化を要する語に対する PCMA の正答率評価を想定し、そのなかで辞書符号付けに対する評価は PCMA の性能を確認する参照点として併せて試行した。研究実施中、PCMA の選択した索引用語と出現した傷病名が完全一致しているにもかかわらず確定符号と一致しない事例から、標準病名マスターと患者調査の符号付け基準の符号化方針差が判明した。このため評価分母を二段構成とし、符号化方針差事例を主指標対象には含むが参考指標対象からは除外した。

（本研究の限界）

本研究の限界として、以下の 5 点を挙げる。

第一に、CMA-ES の目的関数と最終評価指標がいずれも令和 5 年患者調査の同一データに基づくものであり、hold-out 評価および過年度データを用いた外的妥当性検証は今後の課題である。

第二に、参照基準が本研究の対象とする傷病名符号化プロセスの出力そのものであり、本研究の正答率は再現性指標であって、独立した医学的正解との一致率ではない。

第三に、参考指標対象（110,670 語）の追加除外項目は研究結果から事後的に同定したカテゴリであり、研究者の解釈に依存する自由度を含む。主指標は 85.09%（111,984 語）に固定し、参考指標は感度分析の参考値として扱う。

第四に、形態素解析との直接比較は両者の第 1 位候補同士の対比であり、形態素解析が候補を提示した語に対する手付けの最終結果との比較ではない。

第五に、評価単位は得られた種類数であり、患者調査の集計（延べ出現数）への直接的影響を表現するものではない。延べ出現数による加重正答率の評価は今後の課題である。

（PCMA の試行的工夫の到達と残存する構造的課題）

本研究では、文字列照合の限界に対応するため、複合的な近似評価、候補の複数提示、照合と概念点検の分離、ならびに複数部位記載に対する多部位対応の符号への置換を試行的に導入した。これらの工夫により、出現した傷病名に対して補助的判定根拠を提示する基盤を試作した。一方で、標準病名マスターを照合源とする方法には構造的課題が残る。第一に、標準病名マスターに ICD-10 コードが付与されていない病名が存在する。第二に、修飾語付加により ICD-10 コードの妥当性が変動する。第三に、標準病名マスターと患者調査の符号化方針には差異がある。これらの構造的課題は、本研究の試行的工夫では完全には解消されない。

（今後の方向性）

今後の方向性として、(1) 診療情報管理士複数名による独立判定と評価者間一致度に基づく独立参照基準の構築、(2) 過年度データによる外的妥当性検証、(3) 標準病名マスターと患者調査の符号付け基準間の対応関係整備による符号化方針差の構造的理解、(4) 編集距離を考慮した近似照合による未登録誤字・略語への対応強化、(5) 形態素解析が候補を提示しなかった語のうち PCMA も候補を提示しなかった語に対する代替的照合手段の検討、を挙げる。

E. 結論

令和 5 年患者調査の主指標対象（111,984 語）における PCMA 第 1 位候補と確定符号との ICD-10 における 3 桁分類以上での一致率は 85.09%であった。辞書符号付けでの正答率は 92.03%、辞書照合で処理されなかった語（75,101 語）における PCMA の正答率は 81.68%であった。このうち形態素解析が候補を提示し PCMA と直接比較可能であった 59,375 語では、PCMA の正答率が形態素解析を 11.47 ポイント上回った（p<0.0001）。手付け符号化における判断負荷を軽減する可能性がある語（75,101 語）を 3 層に層別化し、形態素解析と PCMA が一致した符号を提示する第 1 層（41,754 語、当該層正答率 93.03%）を識別した。

以上より、PCMA は、傷病名符号化の最適化（正確性の担保と判定根拠の明示性・再現性）および効率化（辞書照合で処理されなかった語、とくに形態素解析が手がかりを与えなかった語に対する手付け作業の調査・探索負担の軽減可能性）に資する並行照合の手段として機能し得る。本指標は確定符

号との一致率であり、独立参照基準の構築および過年度データによる外的妥当性検証は今後の課題である。

F. 健康危険情報

特になし。

G. 研究発表

1）論文発表

学術誌投稿準備中。

2）学会発表

該当なし。

H. 知的財産権の出願・登録状況

特許取得、実用新案登録、その他いずれも該当なし。

I. 参考文献

1) World Health Organization. International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10). Geneva: WHO; 2019.

2) 厚生労働省政策統括官付参事官付保健統計室. 患者調査の概況(令和 5 年). 厚生労働省; 2024.

3) 一般財団法人医療情報システム開発センター (MEDIS-DC) . ICD10 対応標準病名マスター Ver.5.17. 2024.

4) 厚生労働省. 疾病、傷害及び死因の統計分類 (ICD-10 (2013 年版)) 内容例示表. 厚生労働省; 2016.

5) Hansen N, Ostermeier A. Completely derandomized self-adaptation in evolution strategies. Evolutionary Computation. 2001;9(2):159–195.

6) Hansen N. The CMA evolution strategy: a tutorial. arXiv:1604.00772. 2023.

表 1．主指標対象における傷病名符号化工程別の内訳

工程	語数（n）	割合（%）	工程の内容
辞書符号付け	36,883	32.94	傷病名辞書との照合により符号が付与された語
形態素解析が候補を提示した語に対する手付け	59,408	53.05	辞書照合で不一致となった語のうち、形態素解析により候補が提示された語に対する手付け
手付け符号化	15,693	14.01	形態素解析により候補が提示されなかった語に対する用手的な手付け符号化
合計	**111,984**	**100.00**	

注：主指標対象（111,984 語）＝ 傷病名（112,133 語）− データ例外（14 語）− 非傷病名（135 語）。

表 2. 評価上の除外項目

区分	項目名	語数（n）	内容
主指標対象除外	データ例外	14	記号・数字・空文字のみで、傷病名として成立しない対象
主指標対象除外	非傷病名	135	修飾語のみ、部位のみ、診療科名、後置語、なし表現などで構成され、傷病名として成立しない語
参考指標対象 追加除外	ICD 未設定または候補なし	116	標準病名マスターに収載されるが、当該病名単独では ICD-10 コードが一意に定まらない傷病、または PCMA の照合で有効な候補が選定されなかった対象
参考指標対象 追加除外	確定符号不明系	15	患者調査側の確定符号が確認できない対象（確定符号不明 9 ＋ 索引一致・確定符号不明 6）
参考指標対象 追加除外	符号付け方針の差異	1,184	PCMA が選択した索引用語と病名が完全一致するが、標準病名マスターの ICD-10 コードと患者調査側の確定符号が一致しない対象（索引一致・確定符号相違 746 ＋ 左右不影響・確定符号相違 438）

注：主指標対象は、傷病名 112,133 語から、データ例外 14 語および非傷病名 135 語を除外した 111,984 語である。参考指標対象は、主指標対象 111,984 語から、ICD 未設定または候補なし、確定符号不明系、符号付け方針の差異に該当する語を除外した 110,670 語である。なお、ICD 未設定または候補なしと確定符号不明系の間に 1 語の重複があるため、各除外項目の単純合計を減じた値（110,669）とは 1 語の差がある。ICD 未設定または候補なし 116 語の符号化工程別内訳は、辞書符号付け 28 語、形態素解析を用いた手付け符号付け 47 語、手付け符号化（担当者による符号付け）41 語であった。

表 3. PCMA の処理段階と役割

段階	処理内容	出力
① 正規化	全角半角統一、小書き文字統一、 句読点・空白除去、修飾語識別	正規化された病名
② 照合	標準病名マスターの全索引用語との 文字類似度を PCMI として算出	候補リスト （PCMI が高い順、上位第 5 位候補）
③ 概念点検	悪性／良性、両側／単側、急性／慢性、 章・サブ分類等の医学概念を確認し、 不整合候補を抑制	PCMA 第 1 位候補

表 4. PCMI の主要な構成要素と算出例

記号	観点	算出例（「胃癌」vs「胃悪性腫瘍」）
s_base	共通文字率	「胃」が共通 → 1/2 ＝ 0.50
p_diff	非共通ペナルティ	候補側に「悪」「性」「腫」「瘍」 → ペナルティ加算
s_order	語順整合度	「胃」が両者の先頭 → 高値
s_len	索引用語長係数	候補語長＞入力長 → 適切と判定
s_gram	2-gram 一致率	共通の 2 文字連続なし → 低値
s_morph	形態素類似度	「胃」「癌」／「胃」「悪性」「腫瘍」 → 「胃」が一致
s_mod_coh	修飾語タイプ整合度	両者の修飾語（部位・側性・重症度等）タイプの整合度を評価
s_core	修飾語除去後文字集合共通率	傷病名から修飾語を除去した部分と候補索引用語との文字集合共通率を評価

表 5. 工程別および全体正答率

集計区分	語数（n）	正答数	正答率（%）	95%CI
辞書符号付け	36,883	33,945	92.03	91.75–92.31
形態素解析が候補を提示した語に対する手付け	59,408	49,594	83.48	83.18–83.78
手付け符号化	15,693	11,749	74.87	74.18–75.54
主指標対象 全体	**111,984**	**95,288**	**85.09**	**84.88–85.30**
参考指標対象	110,670	95,288	86.10	85.90–86.30

注：参考指標対象は感度分析の参考値（除外項目の詳細は表 2）。

表 6. PCMA と形態素解析の出力 ICD-10 コード（ICD-10 の 3 桁分類単位）と確定符号一致のクロス集計（n＝59,375）

出力 ICD-10 コードの関係	語数（n）	割合（%）	両方とも確定符号と一致	PCMA のみ確定符号と一致	形態素解析のみ確定符号と一致	両方とも確定符号と不一致
PCMA と形態素解析が同一の 3 桁分類を提示	41,754	70.32	38,844	—	—	2,910
PCMA と形態素解析が異なる 3 桁分類を提示	17,621	29.68	—	10,731	3,920	2,970
合計	**59,375**	**100.00**	**38,844**	**10,731**	**3,920**	**5,880**

注：PCMA 正答率 49,575／59,375 ＝ 83.49%、形態素解析正答率 42,764／59,375 ＝ 72.02%、差 ＋11.47 ポイント（McNemar 検定 χ^2＝3,165.39、p<0.0001）。

表 7. 辞書照合で処理されなかった語 75,101 語における 3 層構造

層	条件	語数（n）	辞書照合で処理されなかった語 75,101 語に占める割合（%）	層の役割
第 1 層	形態素解析と PCMA が一致した符号を提示	41,754	55.60	異なるアプローチが同一の符号に到達
第 2 層	形態素解析と PCMA が一致しないがいずれも符号を提示（PCMI が文字列上の近似度）	17,621	23.46	両者を手がかりに手付け
第 3 層	形態素解析が候補を提示しなかった語に対する手付け符号化対象（PCMA 候補提示あり 15,652／提示なし 41）	15,693	20.89	手がかりがなかった語に対する PCMA 候補提示の可否を整理
3 層化対象計	**—**	**75,068**	**99.95**	**—**

注：辞書照合で処理されなかった語は 75,101 語（形態素解析が候補を提示した語に対する手付けの対象 59,408＋同候補が提示されなかった語に対する手付けの対象 15,693）。辞書符号付けは傷病名辞書により確定符号が付与された範囲であり、当該語に含まれない。「異なるアプローチ」は形態素解析（言語処理）と PCMA（標準病名マスター照合）の対比を指す。第 1 層と第 2 層の合計 59,375 は形態素解析側で第 1 位候補が出力された語数であり、75,068 と 75,101 の差 33 語（第 1 位候補が出力されなかった語）は 3 層化の対象から除外した。各層の PCMA 第 1 位候補正答率は、第 1 層 93.03%（95%CI：92.78–93.27）、第 2 層 60.90%（同：60.18–61.62）、第 3 層 74.87%（同：74.18–75.54）であった。

表 8. 第 1 層 41,754 語における不正答 2,910 語の内訳

区分	語数（n）	割合（%）
アルゴリズム改善余地（残余として算出）	2,171	74.60
符号付け方針の差異に起因	696	23.92
ICD 未設定	28	0.96
確定符号不明系	15	0.52
合計	**2,910**	**100.00**

注：第 1 層は形態素解析と PCMA が一致した符号を提示した語（41,754 語、表 7）。「アルゴリズム改善余地」は不正答全体（2,910 語）から、符号付け方針の差異に起因する語、ICD 未設定、確定符号不明系を除いた残余として算出した。残余として整理された語については、今後のアルゴリズム改善余地として検討する必要がある。

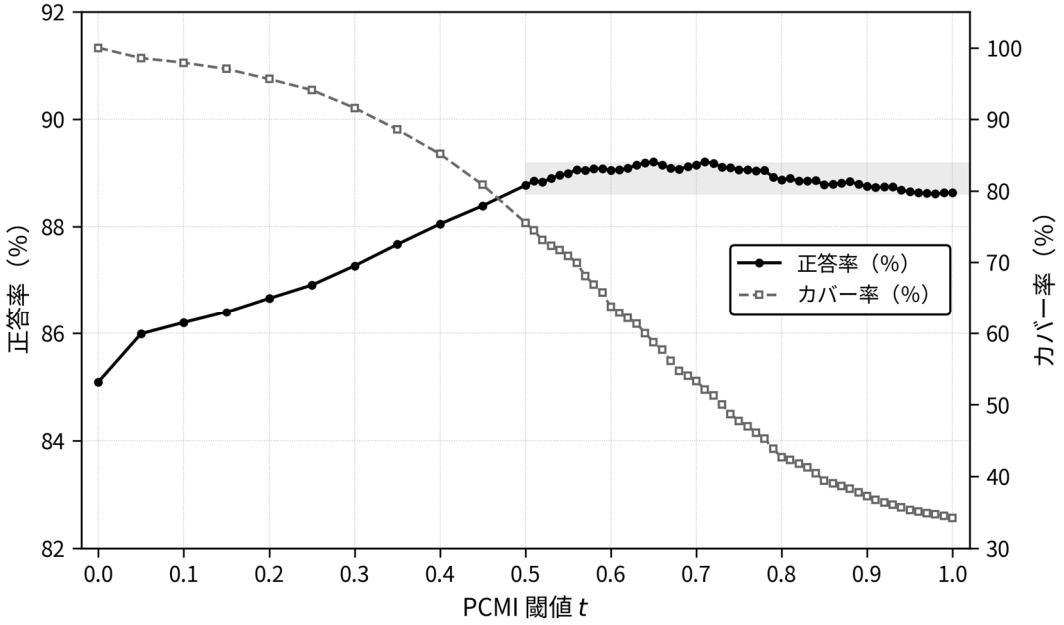


図 1. PCMI 閾値別の正答率とカバー率

注：横軸は PCMI 閾値 t、左縦軸は正答率（%）、右縦軸はカバー率（PCMI≥t の語数比率、%）。閾値を高めても、正答率は 88.6〜89.2%の狭い帯域に留まる。PCMI は候補順位付けを目的として設計された指標である。