2E3-03

臨床医学分野における用語概念間の関係情報を用いた自動 ICD コーディングに関する研究

Automated ICD-coding Using Semantic Relationships in Formal Representations of Medical Concepts

今井 健[*1] 荒牧英治[*1] 梶野正幸[*2] 美代賢吾[*1] 大江和彦[*3]
Takeshi IMAI, Eiji ARAMAKI, Masayuki KAJINO, Kengo MIYO, Kazuhiko OHE

[*1] 東京大学医学部附属病院 Department of Planning, Information and Management, The University of Tokyo Hospital
[*2] 臨床医学オントロジー研究会 Japan Research Group for Medical Ontology
[*3] 東京大学大学院医学系研究科 Department of Medical Informatics, Graduate School of Medicine, The University of Tokyo

The main objective is to create an ontology-based coding support tool for the International Classification of Diseases (ICD-10). There were example-based and knowledge-based approaches for automatic coding tasks in the preceding studies, however, example-based approach does not have explanation capability for coding result. In this study, we utilized knowledge-based approach using ICD Ontology, which is based on formal representation of disease concepts of ICD-10 categories. In the ICD Ontology, a concept and its label are separately defined, and automatic ICD coding was performed using coding principle and string matching between disease name and concept labels. With only labels in the original ICD Ontology and simple string matching, the coding possibility was only about 30%, but there was remarkable improvement in coding possibility with adding labels and coding principle, which demonstrates the basic feasibility of our approach.

1. はじめに

診療現場における IT 化の推進に伴い、電子的に蓄積された診療情報が日々増加している。これには検査結果などの画像・数値データだけでなく、診断報告書や電子カルテで入力される自由記述形式のテキストデータも含まれる。

近年このようなテキストデータに対し、「特定の特徴を持つ患者情報の検索」や、統計的な処理による「臨床上有用な相関を持つ特徴同士の抽出」や「経営判断のための指標」の導出、あるいは「疾患や症状から国際的な疾病分類体系である International Statistical Classification of Diseases and Related Health Problem (以下 ICD-10) への自動コーディング」を行う、などの知的処理を行う需要が高まっている。

特に現在、診療情報管理士の人手作業による ICD コーディングチェックは膨大なコストがかかっており、自動化手法が望まれているが、これを計算機が行うためには「疾患が発生する部位、特徴、原因、随伴症状」といった知識が必要となる。例えば「"胃癌" とは "胃" に発生する "悪性新生物" である」、あるいは「"筋強直性ジストロフィー" は "筋力低下", "筋萎縮", "脱毛", "白内障", "内分泌異常",.. といった症状を伴う」といった用語概念間の意味関係を計算機が扱えるようにしなければならない。

このような背景のもと、現在我が国において臨床医学分野の用語概念間の関係を記述したデータベース (以下オントロジー) が注目されている。欧米では既に臨床医学分野のオントロジーもしくはそれに近い知識リソースとして The Foundation Model of Anatomy Ontology (FMA), Systematized Nomenclature of Medicine –Clinical Terms (SNOMED-CT) などが存在している。しかし、これらには日本語医学用語が概念の表記ラベルとして含まれておらず、そのまま我が国の診療情報と組み合わせて用いることができない。従って、日本語臨床医学オントロジーの構築が必要で、現在我々もこれに取り組んでいる。

連絡先:今井 健, 東京大学医学部附属病院企画情報運営部
〒113-8655 東京都文京区本郷 7-3-1,
Tel: 03-3815-5411, Email: ken@hcc.h.u-tokyo.ac.jp

2. ICD Ontology

臨床医学オントロジーにて対象とする概念は疾患・所見、解剖学的部位、検査・手技、薬剤など多岐に渡るが、とりわけ疾患概念はその中核を成すもので、膨大な数の概念定義をいかに効率的に収集するかが大きな鍵となる。そこで我々はこれまで疾患オントロジーの構築を目指し、その基盤となるリソースとして ICD-10 分類情報の構造化を行ってきた[今井 07]。疾患の分類体系は、疾患概念を他と区分する情報が最も簡単なものから記述されており、基礎的な意味関係の収集に適していると考えられるからである。

本研究では、これらの作業を通じて蓄積された「用語概念間の意味関係情報に基づく疾患概念定義」を用い、自動 ICD コーディングツールへ応用することを目的とする。

図 1 に ICD-10 の分類情報の一部を示す。ICD-10 は全 22 章に渡って構成され、各章がそれぞれ図のように階層構造になっている。診療情報管理士は、この分類情報を基に、例えば「心臓横紋筋肉腫」という病名を C38.0 にコーディングしている。

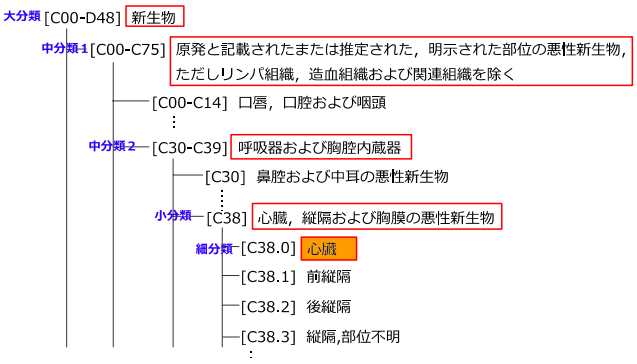

図 1.ICD- 10 の分類階層構造と、分類情報の例

この ICD-10 分類情報の構造化は、各分類項目から文字列ベースで概念を切り出し、<主病態>, <発生部位>, <原因>, <呈する症状>, <障害発生機能>, <随伴病態>, .. など計 143 種類の意味関係を人手付与することで行った。現在のところ「精神系

の障害」や「特殊目的用コード」などを除いた、ICD-10 の主要な 15 章について、作業が完了している。詳細は [今井 07] を参照されたい。以降、この ICD 分類情報に基づいて作成された概念定義データベースを ICD ONTO と呼ぶ。

ICD ONTO は、概念定義の粒度も粗く、オントロジーとしての完成度はまだ低い。一種のライトウェイトオントロジーと位置づけられるだろうが、下記が特徴である。

- 広範囲な疾患概念の基礎的な意味関係を持っている
- 自然言語処理アプリケーションとの親和性を保つため、概念と、それに対する表記ラベルを分けて持っている

今後、これをベースに、トップオントロジーとの接続やヘヴィオントロジーへの深化を進めるという方針である。

2.1 ICD ONTO 中の概念定義

図 2 に ICD ONTO 中の概念定義の具体例を示す。ICD の分類階層における、「大分類」「中分類」「小分類」「細分類」、またそれに含まれる「例示項目」には、FCR(Formal Concept Representation) という概念定義が記述されている。図 2 は第 04 章 E009 の例示項目が持っている概念定義(FCR)である。各行には、この疾患概念が持つ「意味関係」と、各「構成要素概念」、またその概念の「表記上のラベル」と「Cardinality 情報」が記述されている。これらの各構成要素概念同士はフラットではなく、「係り先」情報に基づいて、実際には木構造を成している。

また、ICD ONTO 全体では 30,127 個の FCR と、65,776 個の構成要素概念が含まれている。

図 2.ICD ONTO 中の概念定義 と、構成要素概念の表記ラベル

3. ICD ONTO を用いた ICD コーディング

3.1 従来の ICD コーディング支援研究

自動 ICD コーディングに関する研究は歴史が古く、90 年代から多くの試みが成されてきたが、未だ解決されていない。これらの先行研究は大きく分けて 2 つに区分できる。

- ICD コードに対応する診療文書や病名を用例として集め、それとの文字列ベースの類似度による手法 (用例ベース)
- ICD 分類情報自体を構造化し、コーディング知識を記述したリソースを用いた手法 (知識記述ベース)

前者の例として、[荒牧 07], [Tagliabue 06], [Pakhomov 06], [Michel 95] などがあり、後者の例として、[He'ja 07], [Fabry 03], [Bouchet 98], [Bernauer 98], [Delamarre 95] などがある。前者の方は実装が簡単であるという利点があるが、高精度なものはない。精度向上のためには用例を大量に追加しなければならないが、各コードに対する用例を大量かつ均等に収集することは困難であることと、出力結果に対する判断理由が得られないことが欠点である。一方で、後者は、出力結果が得られた理由を示すことが可能であることが利点だが、記述しなければならない知識が膨大で、ほとんどの研究はモデル提案か、限定した章に対する検討にとどまっている。

本研究は後者の手法に分類されるが、例えば[He'ja 07] は欧米での臨床医学オントロジーの 1 つである Galen を用いて ICD 分類情報の概念定義を記述しており、本研究に非常に近いものである。しかし、2 つの章しか対象にしておらず、他の章も同じフレームワークで記述できるかは不明である。本研究では、ICD の主要な 15 の章全てを記述したリソースを用いており、(1) 網羅性の点で他に類を見ないこと、ならびに (2) 概念定義と自然言語処理アプリケーションとの親和性・整合性を保つために、概念と表記ラベルを別々に取り扱っていること、が他の類似研究と大きく異なる点である。

3.2 ICD ONTO を用いた ICD コーディング手法の概略

本手法で行う ICD コーディングは、2 つの概念に対しその概念定義における構成要素同士を比較したときに、「対応する構成要素が同じまたは下位であれば、全体としても、下位概念である。」という原理に基づいている。(図 3 の Pattern1 参照)

コーディングシステムに入力された病名は、その部分文字列と、ICD ONTO 中の表記ラベルとのマッチングによって、どれかの FCR に合致するか、その下位概念となったときに、その FCR が持つ ICD コードが付与されることになる。

実際の入力病名(“甲状腺機能亢進症”) をその構成要素文字列に分解する際、”甲状腺 + 機能亢進症”, “甲状腺機能 + 亢進症” など複数の分割方法があるばかりでなく、それぞれの文字列は ICD ONTO 中のあちこちで表記ラベルとして使われているため、どの構成要素概念と対応するのかは、多くの可能性がある。システムは、これら全ての可能性を考え、同一 FCR 由来の表記ラベルで病名が構成されるかを探索する。

4. コーディング実験

4.1 実験 1 (ICD ONTO のみ)

ICD ONTO における概念定義が ICD Coding にどの程度応用可能であるのかの調査実験を行った。

材料として、我が国における「標準的な病名」と「対応する ICD コード」を収載したデータベースである「標準病名マスター[1]」を用いた。これに収載されている病名のうち、ICD ONTO を作成した主要な 15 の章に関係する 34,484 病名を対象とした。

(Step1) ICD ONTO 中の疾患概念定義における「構成要素の表記ラベル」を全て抽出した辞書を作成、そのリソース (10,497 語) のみを用いて入力病名からの構成要素解析を行った。Tagger は自作ツール YOMOGI を用い、構成要素分割結果は、スコアに基づき 10-best 解を用いた。
出力される、各「病名構成要素」は「ICD ONTO 中で、どの概念定義(FCR)中のどの構成要素の表記ラベルであるか」という情報を持っている。

(Step2) 入力病名の「構成要素分割結果の 10-best 解」に対し、順に「Coding 可能性」をチェックした。どれかの分割結果で下記の条件を満たせば「Coding 可能」と判断される。

(Coding 可能条件)
「病名の構成要素『全て』が、正解 ICD コード or 上位分類の FCR が持つ『構成要素概念』に対するラベルである」

「Coding 可能」を評価基準にするのは、複数の ICD 候補が得られた際、それらから最終解を決定するためにはもっと高度な知識が必要であり、現在はまだそれを実装していないからである。
また、実験 1 は FCR に対するシンプルな一致検索である。
3.2 節で述べたコーディング原理は、実験 2 にて使用する。

[1] http://www.dis.h.u-tokyo.ac.jp/byomei/

4.2 実験 2 (ICD ONTO + 追加ラベル)

次に、実験 1 で Coding 不可となった病名だけを対象に、追加実験を行った。元々の ICD ONTO だけでは構成要素概念と対応づけするための表記ラベルが十分ではないため、「標準病名の部分文字列から、各構成要素概念の表記ラベルとして適切なもの」を収集し、ICD ONTO に追加した。この際、追加で収集した表記ラベルは、下記の 2 種類である。

- (A) 構成要素概念のラベル
- (B) 構成要素概念の「下位概念」に対するラベル

例えば、(A)「<呈する症状・所見> 顆粒状変性」 のラベルとして、"顆粒性"、(B) 「<原因> 外的因子」という構成要素概念の「下位概念」に対するラベルとして "放射線性" などである。
これに伴い、Coding 可能条件は以下のように変更した。これは 3.2 節のコーディング原理に基づくものである。

(Coding 可能条件)

「病名の構成要素『全て』が、正解 ICD コード or 上位分類の FCR が持つ『構成要素概念あるいはその下位概念』に対するラベルである」

5. 結果

5.1 実験 1 (ICDONTO のみ)

システムが「Coding 可能性あり」と判断した病名数とその割合 (Coding 可能率) を以下表 1 に示す。[Coding 可能] 欄の区分は以下の通りである。

- [D] ある分類項目の FCR と直接対応したもの
- [R1] 上位分類の構成要素概念も用いることで はじめてコーディングできるもの

章	標準病名数	Coding 可能		Coding 可能率
		D	R1	
01	2,982	1,041	43	36.4 %
02	3,648	295	536	22.8 %
03	914	290	19	33.8 %
04	1,845	509	16	28.5 %
06	1,854	418	29	24.1 %
07	1,768	513	40	31.3 %
08	488	160	8	34.4 %
09	1,706	629	52	39.9 %
10	1,036	340	35	36.2 %
11	2,856	808	59	30.4 %
12	1,270	390	65	35.8 %
13	2,038	250	18	13.2 %
14	1,367	437	23	33.7 %
17	2,677	1,065	40	41.3 %
19	8,035	791	237	12.8 %

表 1. ICD-Coding 可能率 (病名⇔ICDONTO)

5.2 実験 2 (ICDONTO + 追加ラベル)

追加実験の結果を表 2 に示す。現在この「表記ラベルの収集」は未だ作業中であり、全ての章に渡って完了してはいない。しかし、これによって得られた結果を ICD ONTO にマージすることで、どのくらいの効果があるのかを見積もるため、現在精査作業が終了しているものだけを用いて行った。

表中の [実験 1 不可] は、実験 1 で Coding 不可能であったものの総数である。この中から一部を選択し([対象]欄)、この中から収集することができた追加ラベルを ICD ONTO のラベルに加えて、その[対象]の病名のみのコーディングを行った。

[Coding 可能] の区分は以下の通りである。

- [D] ある分類項目の FCR と直接対応したもの
- [R2] 上位分類の構成要素概念 や 構成要素概念の「下位概念」に対するラベルを用いることで はじめてコーディングできるもの

また、[Coding 可能率] の区分は以下の通りである。

- [新規] 今回対象とした病名のみに対する結果
- [全体] 全ての [実験 1 不可] 病名に対して、[新規] と同様の割合でコーディングできたと仮定した場合の最終的な「全体 Coding 可能率」の見積もり

章	実験 1 不可	対象	Coding 可能		Coding 可能率	
			D	R2.	新規	全体*
01	1,898	544	137	135	50.0 %	68.2 %
03	605	605	312	23	55.4 %	70.5 %
11	1,989	221	78	82	72.3 %	80.7 %
13	1,770	1,770	435	1,117	87.6 %	89.2 %
19	7,007	1,578	194	1,296	94.4 %	95.6 %

表 2.ICD-Coding 可能率 (病名 ⇔ ICDONTO + 追加ラベル)

6. 考察と今後の課題

6.1 Coding 可能率の向上について

実験 1 の結果では、Coding 可能率はほぼ 3 割前後にとどまっているが、実験 2 では追加ラベルとコーディング原理を用いることで、対象とした範囲全てについて Coding 可能率が劇的に改善している。特に 13,19 章では[D]に比べ[R2]が大幅に多いが、これは FCR とのシンプルな完全一致検索では足りず、コーディング原理が有効に機能していることを示している。実験 2 は、対象が限定的で、対象外の章についての効果は不明であるが、今後他の章についても同様の手法を適用することで、全体的な Coding 可能性の飛躍的な向上が見込まれる。

また本研究では、標準病名から収集した追加ラベルを ICD ONTO に組み込んで、再度標準病名に対してコーディング実験をしているため、完全な評価ではない。この結果から分かることは「ICD ONTO が現時点で標準病名に関する知識をどの程度情報を蓄えられたか」である。今後、全国の病院で入力された標準病名以外の病名を対象とした精度評価を行う予定であり、現在正解付きコーパスを作成中である。

6.2 実験 2 でもコーディングできない事例の分析

実験 2 においてもコーディング出来ずに残った事例は主に 2 つに分類される。

(1) 対応する FCR の構成概念の表記ラベルが部分文字列から得られない場合

例えば、D688「その他の明示された凝固障害」にコーディングされる病名「フィブリノゲン欠乏症」の部分文字列から、

<+主病態> 機能障害
<+障害発生機能> 血液凝固

という「D688 の FCR」の構成概念のラベルは得られない。
「"【人名】＋症候群"」のような病名についても同様である。これらは部分文字列を用いて概念定義とのマッチングを行う手法の限界であり、今後別リソースから「フィブリノゲン欠乏症」に関する不足知識を収集する必要がある。

(2) 本手法における「コーディング原則」に合致しない場合

一般に、「概念 B が概念 A の下位概念」ということは、2 つの概念定義の構成要素を見比べたときに、図 3 における Pattern1, 2 のようになっていること (あるいはその組み合わせ) と考えることができる。つまり、「対応する構成要素が下位概念になっている」or「新しく構成要素が追加された」である。しかし、本手法ではコーディング可能と判断する際の原則として Pattern1 のみを採用し、それ以外のパターンは全てコーディング不可とした。なぜなら、Pattern3 のような事例が発見され、Pattern2 と合わせて詳細な分析が必要、と考えたからである。

Pattern3 を「概念同士の内包的定義に基づく IS-A 関係」と考えるのは抵抗があるが、ICD 分類上は「下位」である。また、Pattern3 は Pattern2 と同様に、「新しく構成要素が追加された」形と見ることもでき、違いは「概念定義の木構造中でどこに構成要素が追加されたか」である。しかし、Pattern3 は、何を主病態と見て概念定義の木構造を形成するか、によって、簡単に Pattern2 のように見かけ上変形することができるため、これは本質的な違いと言えない可能性がある。

今回の実験 2 で用いた「コーディング原理」は、以上のような理由から、安全に概念同士の IS-A が成立すると考えられる Pattern1 のみを用いた訳であるが、今後 Pattern2, 3 を合わせて、「概念同士の IS-A である」あるいは「IS-A ではないが、ICD 分類上は下位である」となる事例の詳細な分析が必要であろう。

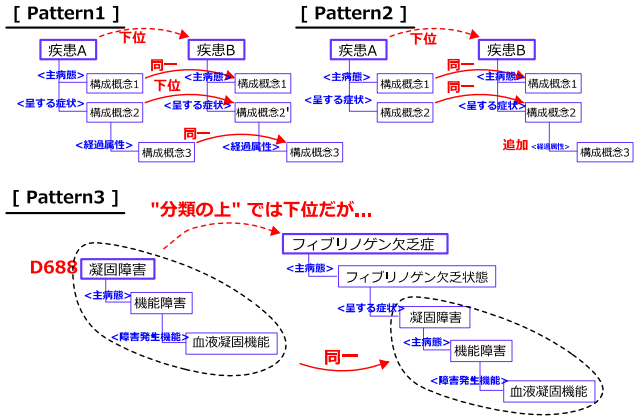

図 3.2 つの概念間の上位下位関係

6.3 最終的な ICD コード決定に向けて

本研究では「Coding 可能率」を用いた評価を行った。これはシステムが出力する「ICD コード候補」の中に正解が入り得るか、を評価するものであり、仮にそれが複数あった場合は次段階の処理が必要となる。特に ICD 分類体系における「その他の～」「詳細不明の～」という見出しを持つコードへの対応づけは大変難しい。例えば M00 以下の分類区分を見てみると、

- M00 化膿性関節炎
  M000 ブドウ球菌性(多発性)関節炎
  M001 肺炎球菌性(多発性)関節炎
  M002 その他のレンサ球菌性(多発性)関節炎
  M008 その他の明示された病原体による(多発性)関節炎

とあるが、ここでもし「インフルエンザ菌性関節炎 (M008)」という病名が入力されたら、「インフルエンザ菌」は「ブドウ球菌」でも「肺炎球菌」ではなく、「その他のレンサ球菌」でもないことをチェックして初めて M008 にコーディングできる。これは表記ラベルマッチングによる消去法では不可能で、「レンサ球菌」にはどんなものがあるのかを予め知らなければならないため、必要な知識量が多い。あえて決定をしないで、複数の候補を提示してユーザに選択を任せるというのも 1 つの方法であるが、一方で診療情報管理士のコーディング支援の観点からは、この決定処理こそが重要である可能性もある。「何が効果的な支援であるか」という点についても、今後より詳細な分析が必要である。

7. おわりに

本稿では、ICD 分類情報から得られた概念定義を用いた、ICD コーディングへの応用手法を提案した。本研究のように ICD の分類情報を元に大規模な概念定義記述を行った例はかつて存在せず、網羅性の観点で他に類を見ない。また、部分的ではあるが、構成要素概念に対する表記ラベルとコーディング原理を用いることで、ICD コーディング可能率が飛躍的に向上することが示された。今後、より一層包括的かつ頑強なコーディング原理を確立すると共に、標準病名以外の自由入力病名に対する評価を行いたい。

参考文献

[今井 07] 今井, 荒牧, 梶野, 美代, 大江: 階層分類情報を用いた疾患オントロジーの半自動構築, 医療情報学, Vol.27 Suppl., pp.700-3, 2007.

[荒牧 07] 荒牧,今井, 梶野, 美代, 大江: 情報検索尺度 Okapi-BM25 と交換可能語ペアを用いた自動 ICD コーディングに関する研究, 医療情報学,Vol.27(1), pp.101-107, 2007.

[Tagliabue 06] Tagliabue G, Maghini A, Fabiano S, Tittarelli A, Frassoldi E, Costa E, Nobile S, Codazzi T, Crosignani P, Tessandori R, Contiero P: Consistency and accuracy of diagnostic cancer codes generated by automated registration: comparison with manual registration, Popul Health Metr, Vol.4, pp.10, 2006.

[Pakhomov 06] Pakhomov SV, Buntrock JD, Chute CG: Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques, J Am Med Inform Assoc, Vol.13(5): pp.516-25, 2006.

[Michel 95] Michel PA, Lovis C, Baud R: LUCID - a semi-automated ICD-9 encoding system, Medinfo, Vol.8 Pt 2, pp.1656, 1995.

[He'ja 07] He'ja G, Surja'n G, Luka'csy G, Pallinger P, Gergely M: GALEN based formal representation of ICD10, Int J Med Inform, Vol.76(2-3), pp.118-23, 2007.

[Fabry 03] Fabry P, Baud R, Ruch P, Le Beux P, Lovis C: A frame-based representation of ICD-10, Stud Health Technol Inform, Vol.95, pp.433-8, 2003.

[Bouchet 98] Bouchet C, Bodenreider O, Kohler F: Integration of the analytical and alphabetical ICD10 in a coding help system - Proposal of a theoretical model for the ICD representation, Medinfo, Vol.9 Pt1, pp.176-9, 1998.

[Bernauer 98] Bernauer J, Schoop D: Formal classification of medical concept descriptions -graph-oriented operators, Methods Inf Med., Vol.37(4-5), pp.510-7, 1998.

[Delamarre 95] Delamarre D, Burgun A, Seka LP, Le Beux P: Automated coding of patient discharge summaries using conceptual graphs, Methods Inf Med, Vol.4(34), pp.345-351, 1995.