様式 C -19

科学研究費補助金研究成果報告書

平成 21 年　5 月 15 日現在

研究種目：　若手研究（B）
研究期間：　2007～2008
課題番号：　19700128
研究課題名（和文）　自由入力病名から国際疾病分類体系へのオントロジーに基づく自動分類手法に関する研究
研究課題名（英文）　Automated Disease Name Classification into International Classification of Diseases based on a Medical Ontology

研究代表者
今井　健（IMAI TAKESHI）
東京大学・医学部附属病院・特任助教
研究者番号　90401075

研究成果の概要：臨床現場において現在人手で行われている「病名の国際疾病分類体系(ICD-10)への分類(コーディング)」を支援するため、(1) ICD10 分類体系の意味構造を「疾患とそれが持つ特性とのリンク構造」として表現したデータベース(ICD オントロジー)を作成し、(2) 任意の自由入力病名と ICD オントロジーとのマッピング手法を開発した。ICD オントロジーは標準病名の約 85%を自動コーディング可能な知識量を網羅し、これにより全国病院の自由入力病名の約 60%が自動コーディング可能であるという結果を得た。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	2,200,000	0	2,200,000
2008 年度	1,100,000	330,000	1,430,000
年度			
年度			
年度			
総　計	3,300,000	330,000	3,630,000

研究分野：　総合領域
科研費の分科・細目：　情報学・知能情報学
キーワード：　臨床医学オントロジー, ICD-10, コーディング支援システム, 複合語解析, 自動分類

１．研究開始当初の背景

(1) ICD コーディングとその現状

現在、我が国の医療現場では医療費削減の観点から、長く「出来高払い」を基本としてきた保健医療制度に対する抜本的な見直しとして、平成 15 年度以降「包括支払い制度」の導入が進められている。この包括支払い制度は診断群分類(DPC)を用いて決定され、この分類を決定する主たる情報が、WHO(世界保健機関)が作成した国際疾病分類体系である ICD-10(International Statistical Classification of Diseases and Related Health Problems Tenth Revision) である。そのため、病名に対し ICD-10 分類コードを付与する作業(ICD コーディング)は非常に重要なタスクであるが、現在は診療情報管理士の人手作業で行われている。

我が国では、ICD コードに対応した標準病名マスターの開発が進められてきたが、標準化は未だ十分ではない。実際の医療現場では長い複合語で構成される病名のバリエーションが日々大量に生成されており、

標準病名マスターだけを用いて ICD コーディングを行えるのは全体の 45% 程度である。また、その病名表現の多様性のため人手作業であっても ICD コーディングは難しく、3 人の診療情報管理士の一致率も 54% 程度にとどまる。そのため誤ったコーディングにより実際よりも高い/安い診療報酬が請求される問題、保健医療政策決定上必要な統計データの算出基準の信頼性の問題が指摘されている。加えて、診療情報管理士の人手も大幅に不足しており、計算機による ICD コーディング支援手法が希求されている。

(2) 国内外の研究動向と要請事項

自動 ICD コーディングを目指した研究は国内外にごく少数しか存在せず、従来 ICD コードが既知である病名と文字列の類似度を用いてマッピングする手法などがあるが、実装が簡単なものの精度が低いという問題があった。ICD-10 分類体系は、疾患を「主病態」「原因」「発生部位」「臨床上の特徴」など様々な観点からの意味関係を用いて階層的に分類するものであり、精度向上のためには、表層の文字列情報だけでなく、これら疾患概念が持つ「意味関係」を扱う必要があるという指摘がされている。

2．研究の目的

以上の背景のもと本研究は以下の 2 つの要素を基に「任意の病名に対する自動 ICD コーディング手法の確立」を実現しようとするものである。

(1)ICD-10 分類体系の持つ意味構造を表現したオントロジーの構築手法の確立と評価

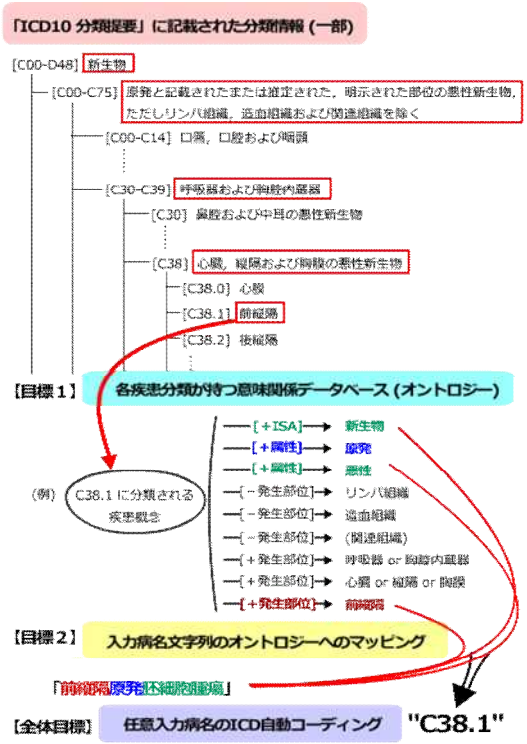
疾患を中心としたあらゆる意味関係を記述したオントロジーを 1 から開発するには膨大なコストを要する。本研究では ICD コーディングという目的に必要なサブセットを効率的に得るため、ICD-10 分類体系自体が持つ情報から「分類に必要十分な意味関係の抽出」を目指す。具体的には各疾患カテゴリが持つ主病態，原因，発生部位，症状/所見，etc.と行った意味関係を整理したデータベース(以下「ICD オントロジー」)を構築する。

(2)入力病名文字列の、上記オントロジーへのマッピング手法の確立と評価

臨床現場では「抗酸菌染色陽性活動性結核性肺線維症」のような長い複合語病名が頻出する。この解析には用語表記の揺れの回収だけでなく、ICD 分類に必要な意味関係に応じて語の分割粒度を動的に決定する必要がある。これは一意的に文字列の分割を決定する形態素解析では不十分であるため本研究では入力病名の分割粒度を可変にした複合語解析手法を確立する。またそれを用いた ICD オントロジーとのマッピングに基づく ICD コーディング手法を確立し、評価する。

3．研究の方法

本研究全体のスキームを以下図 1 に示す。



(図 1 本研究のスキーム)

(1) ICD オントロジーの構築

①プロトタイプオントロジーの構築

ICD10 の分類情報を記載した「ICD10 分類提要第 2 巻 (2003 年度版)」(全 1,070 ページ)を対象にし、「特定のコードに分類される疾患が持つ意味関係」を抽出し、疾患を中心としたプロトタイプ ICD オントロジーを構築した。

例えば、上記図 1 で「C38.1 に分類される疾患概念(例: 前縦隔原発胚細胞腫瘍)」は、ICD の分類情報に書かれている内容を上位階層から順に辿ることで、図に示す複数の意味関係を持っていることが分かる。このような意味関係を ICD10 の主要な 15 章の全疾患カテゴリに対して記述した。意味関係の種類は「主病態」「発生部位」「原因」「属性」「随伴症状」「基礎疾患」など大きく区分されるものでも 15 種類以上存在するが、これらを区別しながら情報付与を行うには ICD 分類に特化した専門知識が必要であるため、診療情報管理士のべ 20 名の協力のもと作業を行い、別の 2 名により精査作業を行った。

②概念と表記ラベルとの切り分け・表記ラベル収集

①で構築したプロトタイプオントロジーは、分類情報に記載されている文字列を可能な限り尊重して構築したが、例えば「<属性> 顆粒性」という意味関係における「顆粒性」という文字列は、本来「<呈する症状/所見>が『顆粒状変性』という[異常状態]である」ということに対する「別名(表記上ラベル)」に過ぎない。このような事例に対し、登録概念としてふさわしいものと、その表記上のラベルとの切り分け作業を行った。

また、同一概念に対する多様な表記ラベルを収集するため、我が国の「標準病名マスター」中の約 25,000 病名の部分文字列から、表記上ラベルの収集を行った。例えば「巣状」という部分文字列は、「<発生部位(範囲)> が『全糸球体の 5 割未満』という概念を表現するラベル」として収集される。

③ 粒度の細かい下位概念の追加

ICD 分類情報だけから得られる意味関係と概念/表記ラベルだけでは臨床現場の多様な病名表記に対応できない。例えば「<発生部位> 下顎骨」に対する「下顎骨骨頭」や、「<原因> 外的因子」に対する「放射線被曝」など、多様な部分/下位概念を追加する必要がある。そこで標準病名マスター中の病名を対象に、より粒度の細かい下位概念を収集し、ICD オントロジーの疾患カテゴリ記述に使われている各構成要素概念の下位概念として拡張した。

これら②,③の作業も①と同様臨床情報管理士の協力により行い、別の 2 名により精査作業を行った。

④ 記述フレームワークの高度化

各疾患カテゴリの意味関係記述フレームワークと、現在のオントロジー工学理論との整合性を取り、洗練されたオントロジーとするため、現在東京大学大江研究室と大阪大学溝口研究室で構築が進められている「臨床医学トップオントロジー」の情報モデルを参考にし、必要に応じて記述フレームワークの修正随時行った。また、現在 WHO(世界保健機構)主導のもと、ICD-10 の意味構造を記述し ICD-11 へ改訂するプロジェクトが世界規模で取り組まれようとしているが、その改訂会議議長である米国 Mayo Clinic の Christopher Chute 教授研究室と連携し、ICD-11 での疾患記述モデルとの親和性を向上するため随時記述フレームワークの修正を行った。

(2) 病名入力文字列からオントロジーへのマッピング

①入力病名解析器の構築

入力病名を構成要素に分割し、ICD オントロジーとマッピングする際、求められる分割粒度はコンテキストに依存し、必ずしも言語学上の形態素とは一致しない。そのため ICD オントロジー中の表記ラベルを辞書とし、一意な形態素解析ではなく N-Best 分割解を提示する入力病名解析器(YOMOGI)を構築した。

②ICD オントロジーとのマッピングによる ICD コーディングツールの実装と評価

①で得られた病名解析器を用い、入力病名の構成要素とオントロジー中の概念をマッピングするツールを構築した。さらに、得られた複数のマッピング候補を元に、最適な ICD コードを選択する ICD コーディングツールを実装した。

また、臨床情報管理士の協力のもと、全国病院の自由入力病名に対して ICD コード付与とその判断理由の記述作業を行い、精査作業の後「明確な理由で ICD コードが一意に決まるもの」だけを選別、結果 1,255 の「病名と対応する ICD10 コード」の正解セットを作成した。またこれらを用いて、自動 ICD コーディングの精度を評価した。

4．研究成果

(1) ICD オントロジー

ICD-10 の主要な 15 章における計 15,466 個の疾患カテゴリに対し、全体でのべ 37,472 個の構成要素概念（異なる構成要素概念は 8,385 個）と、68 種の意味関係(ロール概念)を用いた木構造での概念定義記述データベースが得られた。データベースは XML で記述され、図 2 のようにオントロジーブラウザにて閲覧可能である。

(図 2 ICD オントロジーブラウザ)

(図 3 疾患カテゴリの概念記述)

個々の疾患カテゴリの概念記述は図 3 に示すスロット型の記述方式で、意味関係、木構造を表現する係り先情報、Cardinality 情報、構成要素概念、表記上ラベルらの情報から構成されている。

また、個々の構成要素概念はさらに粒度の

細かい「部分概念・下位概念」と関連づけられていることもあり、それら拡張部分は8,033 個の追加概念と表記ラベルから成る。

(2) 入力病名文字列の ICD オントロジーへのマッピング手法の確立と評価

まず、入力病名文字列は、ICD オントロジー中の計 14,663 個の異なる表記ラベルを辞書とした病名解析器 YOMOGI の 10-Best 分割結果を用いて、構成要素(部分文字列)が抽出される。そして各部分文字列（例:「非中毒性」）は、それが未定義語でない限り、ICD オントロジー中に定義された表記ラベル情報を介して、何らかの疾患カテゴリが持つ構成要素概念(例:「<症状/所見> 甲状腺調節機能正常」)、あるいはその下位概念とマッピングされる。当然、該当する構成要素概念は複数の疾患カテゴリの概念定義に現れることもある。

次に、入力病名は ICD コーディングツールにより、「その各構成要素をマッピングした全ての可能性を考慮した結果、最も被覆度の高い疾患カテゴリの ICD コード」にコーディングされる。また「具体的にどの部分文字列がオントロジー中のどの構成要素概念あるいはその下位概念を指し示すか」という形式で、コーディング結果の理由も出力される。

これらのシステムを用いて、既に ICD コードが既知である標準病名マスター(34,484 病名)と全国病院で入力された自由入力病名から作成した ICD コード付き正解セット(1,255 病名)に対して自動コーディング実験を行い、(A) ICD オントロジーは現在標準病名の 85%を自動コーディング可能な程度の知識量を保有し、(B) その知識を用いることで 60%程度の自由入力病名をコーディング可能であるという結果を得た。

(3) 成果の位置づけ・今後の展望

現在 WHO 主導の元、ICD10 分類情報の構造化記述を目指した ICD11 改訂プロジェクトが行われているが、本研究の ICD オントロジーはその先駆けとし、高い網羅性でそれを実現したものである。特に疾患概念記述モデルの一部は既に ICD11 の情報モデル構築のための有用な参考情報となりつつあり、ICD オントロジーは単にコーディングのための知識にとどまらず、それ自身が教育への応用など多様な可能性を持った貴重なリソースであると位置づけられる。

またそれに基づいた自動コーディング手法は、知識ベースに基づいて ICD コーディングを行う研究としては世界で従来例を見ない規模でこれを実現したものであり、用例との文字列類似度に基づく従来手法に比べても成績の向上のみならず (A)用例の有無に関係なく任意の ICD コードへ分類可能である点、(B)分類理由を出力できる点が優れている。特に(B)は複数の候補が出力された際、後の人手によるスクリーニングの手間を大幅軽減するために重要な特徴である。この成果を元に今後、診療情報管理士のコーディング精度向上や時間の短縮を目指した支援ツール、さらには教育ツールへの応用が見込まれる。

５．主な発表論文等

〔雑誌論文〕（計 6 件） 全て査読有

① Takeshi IMAI, Hiroko KOU, Jun ZHOU, Kouji KOZAKI, Riichiro MIZOGUCHI, Kazuhiko OHE. Japan Medical Ontology Development Project for Advanced Clinical Information Systems. Proceedings of 10th International HL7 Interoperability Conference, pp.42-46, 2009.

② 今井　健, 荒牧英治, 梶野正幸, 美代賢吾, 大江和彦. 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. 人工知能学会 第 22 回全国大会論文集, 2E3-03:1-4, 2008.

③ 国府裕子, 周俊, 古崎晃司, 今井　健, 大江和彦, 溝口利一郎. 臨床医学オントロジーの構築に関する基礎的な考察. 人工知能学会 第 22 回全国大会論文集, 2E3-01:1-4, 2008.

④ Imai T, Aramaki E, Kajino M, Miyo K, Onogi Y, Ohe K. Finding malignant findings from radiological reports using medical attributes and syntactic information. Medinfo 12(Pt.1), pp.540-544, 2007.

⑤ Aramaki E, Imai T, Kajino M, Miyo K, Ohe K. Statistical selector of the best multiple ICD-coding method. Medinfo, 120(Pt1), pp.645-649, 2007.

⑥ 今井　健, 荒牧英治, 梶野正幸, 美代賢吾, 大江和彦. 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 27(Suppl.),pp.700-703, 2007

６．研究組織

(1)研究代表者

今井　健 (IMAI TAKESHI)

東京大学・医学部附属病院・特任助教

研究者番号: 90401075

(2)研究分担者

(3)連携研究者