別添資料２

標準病名マスターを用いた死因病名に対する ICD-10 コーディング実験

概要

死亡個票の「死亡の原因」に含まれるⅠ欄・II 欄病名は自由記載である。これに対し、標準病名マスターを用いて ICD-10 コードを付与する(ICD-10 コーディング) ことができる割合について調査を行った。全てのⅠ欄・II 欄病名が ICD-10 コーディングできれば、その結果を Iris に入力し、仮原死因を確定することができる。また結果を確定原死因と比較することで、病名以外の付帯情報による原死因変更の有無が判明するため、これを機械学習により自動分類するための学習用データを得ることができる。さらに、なるべく多くの事例を ICD-10 コーディング可能であれば、それだけ多くの学習用セットが得られることになる。

実験の結果、標準病名マスターを用いて、全てのⅠ欄・II 欄病名に対しほぼ原記載のまま(Step0)、ICD-10 コーディング可能だったのは約 44%であった。また今後、助詞、接続詞の除去/展開と言い換えなどの文字列処理（Step1）を施すことで約 65%程度まで増加するという感触を得た。

はじめに

本年度、平成 27 年～平成 30 年の死亡票とオンライン申請された死亡個票の調査票情報の結合を行った。結合した情報のことを、以下、突合死亡票 DB と呼ぶ。(明神大也 分担報告書を参照)

なお令和元年度は予備実験として、突合死亡票 DB の冒頭から 10,000 件(以下、これを略して突合死亡票 DB と記載する)についてのみ以下の実験を行なった。

実験方法

<実験の概要>

本研究計画は、死亡個票の「死亡の原因」に含まれるⅠ欄・II 欄病名の全て(以下、死因病名)に付与された ICD10 コードに基づき IRIS を用いて導出した原死因と、死亡票に含まれる確定原死因コードを比較し、合致しない症例すなわち従来人手チェックで原死因コードを決定していた症例の原死因コード付与の支援を目的としている。これらの症例におい

ては、死亡病名欄や何らかの付帯情報（傷病名以外の、手術や解剖所見・備考欄・外因死の追加事項など）（以下、これを付帯情報と呼ぶ)の記載内容が影響していると考えられる。

本実験では、死因病名に ICD10 コードを付与する第一段階として、Step0、すなわち文字列処理をほぼ行わずに ICD10 対応標準病名マスターとの対応に基づいて、ICD10 コードを死因病名に付与できるか、突合死亡票 DB を対象に確認する。これに基づき、今後必要となる文字列処置の詳細を考察する。

<前処理>

まず、前処理として突合死亡票 DB に対して以下を行なった。

(1)表記の統一

カタカナ、数字、アルファベット、記号を半角に統一した。(資料 a:utf_hannkaku.pl)

(2)表記揺れの回収

日本医学会：医学用語管理/付表 1 日本語表記のゆれ
(http://jams.med.or.jp/dic/kanji_variance2.html (2019 年 12 月 26 日閲覧))における「その他の表記法」の記載を全て「本辞典で採用した表記法」に変換した。(資料 b：kakikae_10001.sh)

例外処理は以下の通りである：

(a)「その他の表記法」に複数の単語が並列して記載されていた場合には、すべての単語を、対応する「本辞典で採用した表記法」に変換する。

(b)6.異なった用語のあるものにおいて、用語の意味が「【旧】」のように「【】」を利用して記載されている場合には、「【」から「】」までは除外する。

(c)「本辞典で採用した表記法」に複数の用語が記載されている場合は、今回はまずは便宜的に、死亡者数が多いと考えられる方に限定した。たとえば、「知的障害【小児】、精神遅滞【神経】」は(b)の処理も踏まえて「精神遅滞」のみにした。

(d)「その他の表記法」および「本辞典で採用した表記法」に複数の文字列を意味する「・・」がある場合、それを除外して変換する。たとえば、「・・パシー」を「・・パチー」に変換する。これによりミオパシーがミオパチーに変換される。「・・」を除外して変換すると、仮にテレパシーという記載があったとするテレパチーに変換される。しかし、この変換は標準病名とのマッチに影響しないため問題ないと考えた。

<ICD10 コード>

ICD10 コードと標準病名マスターの対応は、ICD10 対応標準病名マスター ver5.00(https://www2.medis.or.jp/stdcd/byomei/download2019.html
(2020 年 7 月 11 日参照))の病名基本テーブルに基づいた。

実験結果

<結果>

死亡病名欄を認識できた 9,985 症例のうち全ての死因病名に ICD10 コードを付与できた症例は 4,351 症例(43.6%)であった。内訳として、記載された死因病名の個数が 1〜5 個の症例数と全体に占める割合はそれぞれ、3,385 症例(33.9%)、810 症例(8.11%)、134 症例(1.34%)、22 症例(0.22%)、0 症例であった。

また、付帯情報がある症例のみに限定したところ、3,134 症例のうち、全ての死因病名に ICD10 コードを付与できた症例は 1,085 症例(34.6%)であった。内訳として、記載された死因病名の個数が 1〜5 個の症例数と全体に占める割合はそれぞれ、803 症例(8.04%)、224 症例(2.24%)、48 症例(0.48%)、10 症例(0.10%)、0 症例であった。

ICD10 を付与できなかった死因病名について、冒頭から 50 件して分類を行なったところ、以下の表の通りとなった。(区分は別添資料 1 の表 1 に準じている。)

ルール		件数	割合(/50)
Step1) 言語処理			
b	助詞、接続詞の除去/展開	30	
b1	「不明」「不詳」と死因病名に記載されている	5	10%
b2	b1 以外	25	50%
c	言い換え	3	
c1	悪性新生物→がん	0	
c2	＜部位＞原発・＜部位＞(における)続発症 →前後の行にがんの記載があれば癌をつける	0	
c3	続発性→転移性	0	
C4	その他の言い換え(縊頸 →首吊り自殺、Paget →ページェット　など)	3	6%
d	文字列処理	4	8%

Step2) これ以上の処理		13	
2-1	誤植	1	2%
2-2	2-1 以外	12	24%

考察

<突合死亡票 DB の読み込みについて>

死亡病名欄を認識できなかった 15 症例は、死亡病名欄や何らかの付帯情報（傷病名以外の、手術や解剖所見・備考欄・外因死の追加事項など）に含まれたカンマを前処理において半角に変換してしまったことが原因であった。令和 2 年度には前処理から含め整理しなおし、より精度の高い実験を行う予定である。しかしこの前処理の問題は全体に占める割合は 0.15%であり、本報告書の結果と考察に大きな影響を及ぼさないと考えている。

<今後必要となる文字列処理>

助詞、接続詞の除去/展開と言い換えで約 65%の死因記載に ICD10 コードを付与できることが期待された。

しかしこれ以外の残りの約 35%は極めて処理が難しいことが予想された。

たとえば、文字列処理には「喘息重積発作」という記載(標準病名は喘息発作重積(J46)と推定される)や「特発性汎血球減少症」(標準病名は汎血球減少症(D619) と推定される)のように文字列の並べ替えや包含の判断で標準病名に変換できると想定されたものも含まれた。また Step2 を大別すると、「貧血進行・子宮からの出血」「摂食不能、脱水」「誤嚥」という正解となる ICD10 コードの判断が難しいもの、「蘇生後脳症」「癌悪質液」といった端的な概念が ICD10 コードに存在しないのではないかと推定されたものの 2 つに分けられた。これらの機械処理は極めて難しいと考えられた。