厚生労働科学研究費補助金(政策科学総合研究事業（統計情報総合研究事業））
「死因統計の精度及び効率性の向上に資する機械学習の検討に関する研究」
分 担 研 究 報 告 書（令和 2 年度）

死亡に関わる調査票情報提供に基づいた ICD10 コード自動付与ツールの作成

研究分担者　香川璃奈　（筑波大学医学医療系・講師）

研究要旨

我が国において人口動態調査は国勢調査と並ぶ国の基幹統計であり、中でも死因統計は最も重要な情報の一つである。診療報酬請求や現在普及が進む電子カルテでは標準病名の採用が進められているが、人口動態調査の死因は自由入力病名が元となっており完全な自動集計は困難である。

本研究では、死因確定作業において目視確認に回る理由の中でも、死因を ICD-10 コードに変換できないという点に焦点をあて、標準病名と完全一致しない死因の記載を標準病名または ICD-10 コードに変換するルールを作成した。その結果をプログラムに実装し、平成 27 年～令和 2 年の死亡票とオンライン申請された死亡個票を突合した情報に ICD10 コードを自動付与したところ、作成したルールを利用しないときと比較して、死因の記載のすべてに対して ICD10 コードを付与できた件数の割合が 57.3%から 80.06%に増加した。

来年度はこの結果に基づき機械学習手法の開発を行う。

A．研究目的

我が国において人口動態調査は国勢調査と並ぶ国の基幹統計であり、中でも死因統計は最も重要な情報の一つである。診療報酬請求や現在普及が進む電子カルテでは標準病名の採用が進められているが、人口動態調査の死因は自由入力病名が元となっており完全な自動集計は困難である。

我々は令和元年度に平成 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なった。結合した情報のことを、以下、突合死亡票 DB(データ数：5,169,031 件)と呼ぶ。これを利用して、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、また助詞、接続詞の除去/展開と言い換えなどの比較的簡便な文字列処理を施すことで、約 65%のⅠ欄・II 欄病名の自動 ICD10 コーディングが可能であるという感触を得た。さらに、Ⅰ欄・II 欄病名を ICD10 コードに変換できたものは約 9 割であった。しかし上記は、突合死亡票 DB の一部を目視で確認した結果である。

そこで本年度は以下の 3 点を行なった。

(1)上記の結果をより詳細に解析した。これにより、標準病名と完全一致しない死因の記載を標準病名または ICD-10 コードに変換するために有効なルールを作成した。

(2)解析結果に基づき、比較的簡便な文字列処理に基づきⅠ欄・II 欄病名の自動 ICD10 コーディングを行うツールを開発した。

(3)開発ツールを利用して実際に突合死亡票

DB の全ての I 欄・II 欄病名を ICD10 コードに変換した。

この際に、(1)で作成したルールを利用した際と、利用しなかった際とで、ICD10 にコーディングできた件数がどれだけ変化したか確認した。

B．研究方法

(1)I 欄・II 欄病名が標準病名と完全一致しない場合の対応ルールの作成

(1-1)I 欄・II 欄病名と標準病名の対応の抽出

I 欄・II 欄病名のうち、標準病名に完全一致しない病名を自然言語記載病名と呼ぶ。

(1-1-1)自然言語記載病名の書き換えルール作成

漢字の誤記、カタカナの表記ゆれ、表記を漢字、ひらがな、カタカナ、英語の間で任意にすれば標準病名になるパターンに対しては、書き換えルールを作成し適用し ICD10 コードを自動付与できるようにすることで、IRIS による自動確定死因付与をより多くの症例に対して行うことが可能になる。これにより、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)による原死因変更の有無の把握およびそれに基づく予測をより高精度に行うことができる。

しかし、突合死亡票 DB の全データの確認にはコストがかかる。そこで、突合死亡票 DB の冒頭 30,000 件と最後 30,000 件の合計 60,000 件の I(ア)欄に記載された自然言語記載病名を目視で確認し、漢字の誤変換、カタカナの表記ゆれ、表記を漢字、ひらがな、カタカナ、英語の間で任意にすれば標準病名に変換されるルールを作成した。

自然言語記載病名のうち、複数の病名が並記されている記載は確認対象から除外した。

(1-1-2)出現頻度の高い自然言語記載病名に対応する ICD10 コードの探索

(1-1)で作成した変換ルールを適用しても標準病名に変換できない自然言語記載病名のうち、特に出現頻度の高い自然言語記載病名に対して ICD10 コードを付与できれば、IRIS による自動確定死因付与をより多くの症例に対して行うことが可能になる。これにより、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)による原死因変更の有無の把握およびそれに基づく予測をより高精度に行うことができる。

そこで以下の作業を行なった。

まず、突合死亡票 DB の冒頭から 50 万行の中で、標準病名と完全一致しなかった自然言語記載病名の上位 10 種類を特定した。該当する 10 種類の自然言語記載病名の各々に対して、突合死亡票 DB において該当する自然言語記載のみが I(ア)欄に記載されており、その他の自然言語記載病名が存在しない死亡表の確定病名として付与されている ICD10 コードを、該当する自然言語記載病名に付与される ICD10 コードとみなした。

(1-2)死亡表の特徴に基づく表記の前処理ルールの作成

(i)表記の統一

カタカナ、数字とアルファベットを全て半角に統一した。

(ii)死亡表の特徴に基づく表記の前処理

死亡表の病名欄には、ICD10 コードを付与する上では不要な詳細な情報(発症の経緯や、病期分類など)が記載される。そこで、以下の前処理を行った。

「(」から「)」までの文字列は削除した。「(」からはじまり最後まで「)」が出てこない場合にも「(」のあとの文字列は削除した。自然言語病名が「)」からはじまる場合は、「)」のみ削除した。

(iii)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[1]に基づいた。[1]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。

例外処理：

(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。

(iv)複数の病名が併記されている時の病名分割ルール
個別の病名欄に複数の病名が併記されている場合がある。この場合は、全角空白、および、読点により病名が区切られていると解釈した。

(2)自動 ICD10 コーディングツールの作成
(1)の処理を行った上で、自然言語記載病名を ICD10 コードに変換するプログラムを作成した。
自然言語病名と ICD10 対応標準病名マスター ver5.00[2]の標準病名が完全一致した場合に、自然言語病名を ICD10 コードに変換した。

(3)突合死亡票 DB の I 欄・II 欄病名の ICD10 コードへの変換
(1)のルールを適用せずに ICD10 コードへの対応を行う処理、および、(1)のルールを適用したあとに ICD10 コードへの対応を行う処理のそれぞれを実施した。

C．研究結果
(1)I 欄・II 欄病名が標準病名と完全一致しない場合の対応ルールの作成
(1-1)I 欄・II 欄病名と標準病名の対応の抽出
(1-1-1)自然言語記載病名の書き換えルール作成
ユニーク数で 4,481 個(延べ 14,747 個)の I(ア)欄の自然言語記載病名が標準病名と完全一致しなかった。全て目視で確認を行い、140 個の変換ルールを追加した。
変換ルールの一部は以下の通りである。全ての変換ルール、および(1-2)(iii)で作成した変換ルールは別添資料４「記載変換ルール」に示す通りである。

表 1: 変換ルールの例。before 列の文字列を after 列に置換することを意味する

before	after
消化菅	消化管
縊首	首吊り自殺
胆管Ca	胆管癌
Ab血症	ｱﾙﾌﾞﾐﾝ血症

(1-1-2)出現頻度の高い自然言語記載病名に対応する ICD10 コードの探索
得られた自然言語記載病名の上位 10 種類と、対応する ICD10 コードは以下の通りである。
なお、「不詳の内因死」という自然言語記載病名が特定されたが、これは「不詳」(表 2)に内包した。

表 2：出現頻度の高い自然言語記載病名に対応する ICD10 コード。(¥D+)は 1 個以上の任意の数の数字以外の文字が存在するという意味である。

自然言語記載病名	ICD10 コード
不詳	R99
不明	R99
縊死	X70
急性心臓死	I461
慢性心不全急性増悪	I509
低酸素脳症	G931
誤嚥	T179
致死性不整脈	I498
虚血性心不全	I255

なお、この結果を確認したあと、(2)では、変換ルール(別添資料５「未コード化傷病名変換ルール」)を利用した。

(1-2)死亡表の特徴に基づく表記の前処理ルールの作成

B．研究方法に記載の通りのルールを作成した。(iii)の結果は先述の通り別添４に示す通りである。その他のルールの具体例は本統括・分担報告書全体の末尾にある**別添資料「Iris 用入力データ作成プログラム」**を確認されたい。

(2)自動 ICD10 コーディングツールの作成

(1)の結果に基づき、自動 ICD10 コーディングツールを作成した。また (1)のルールを適用せずに ICD10 コードへの対応を行う処理と、(1)のルールを適用したあとに ICD10 コードへの対応を行う処理について結果を比較した。

(3)突合死亡票 DB の I 欄・II 欄病名の ICD10 コードへの変換

(1)のルールを適用せずに ICD10 コードへの対応を行う処理を行ったところ、I 欄・II 欄病名のすべての自然言語記載病名に ICD10 コードに変換できたのは 2,960,455 件(57.27%)であった。

(1)のルールを適用したあとに ICD10 コードへの対応を行う処理を行ったところ、I 欄・II 欄病名のすべての自然言語記載病名に ICD10 コードに変換できたのは 4,138,310 件(80.06%)であった。

(4)倫理面への配慮

本研究では統計法 33 条に基づき申請したデータを利用した。申請の通り、インターネットに繋がない端末上でのみデータの閲覧作業を行うことで個人情報に配慮した。

D．考察

今回の処理を通じて、医師が手書きで記載した死亡診断書に基づき、電子的な死亡票データに変換する作業、すなわちオートコーディングの前段階として行う前処理に関して以下の課題が明らかとなった。

作成した変換ルールは、そもそもの死亡診断書の入力記載が標準化されれば確認と作成の手間を除外できるものである。具体的には、「急性」「慢性」などの修飾語句が付与されていた自然言語記載病名、修飾語句に「（」「）」などの記号を独自に付与している自然言語記載病名、そもそも記載欄や記載内容を間違えているように解釈できた自然言語記載病名(例:「膵臓」)、自然言語記載病名が手書きだからこその誤記や省略表現(例:「下日支」「脳卒中」)などである。医師の死亡診断書記入時における標準病名の利用、あるいはその電子データへの転記時での標準病名への変換が実現すれば、このような入力、また入力に基づく解釈に人手を割く必要がなくなり、働き方改革にも貢献できると考える。

また、将来的に日本全国で死亡表の自然言語病名が ICD10 コードへの変換がオフライン上で安全に行われるようになる将来を見据えて、**(2)自動 ICD10 コーディングツールの作成**で作成したツールの一部の機能を docker コンテナとしても実装できることを確認した。

E．結論

本年度研究では、実際の死亡表に記載された病名を ICD10 コードに変換するツールを作成した。独自の対応ルールを利用することで、I 欄・II 欄病名のすべての自然言語記載病名に ICD10 コードに変換できた件数が全体の 8 割を超えた。来年度はこの結果に基づき機械学習手法の開発を行う。

<参考文献>

[1] http://jams.med.or.jp/dic/kanji_variance2.html (2019 年 12 月 26 日閲覧)

[2] https://www2.medis.or.jp/stdcd/byomei/download2019.html (2019 年 12 月 26 日閲覧)

F．健康危険情報

なし

G．研究発表

なし

H．知的財産権の出願・登録状況

なし