

別添資料4

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、突合死亡表 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、突合死亡表 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表

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、突合死亡表 DB の全データの確認にはコストがかかる。そこで、突合死亡表 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、突合死亡表 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、突合死亡表 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) 突合死亡表 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参考資料 4-1「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
消化管	消化管
縊首	首吊り自殺
胆管Ca	胆管癌
Ab血症	アルブミン血症

(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。(YD+)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、厚労省からいただいた変換ルール(参考資料 4-2「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 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) 突合死亡表 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ら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:「脾臓」)、自然言語記載病名が手書きだからこそ誤記や省略表現(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

参考資料4-1 記載変換ルール

before	after
アテトーゼ	アテトース
アミロイド症	アミロイドーシス
アンギナ	アンギーナ
アンドロジェン	アンドロゲン
いおう	イオウ
イボ	いぼ
インシュリン	インスリン
えい児	嬰兒
エストロジェン	エストロゲン
えら	鰓
えん下	嚥下
え死	壊死
おとがい	頤
おり	檻
カ	蚊
かいせん	疥癬
かい癬	疥癬
かかと	踵
がこう瘡	鷺口瘡
がん	癌
ガン	癌
くし	櫛
グルタミン酸オキサロ酢酸トランスフェラーゼ (GOT)	アスパラギン酸アミノトランスフェラーゼ (AST)
グルタミン酸ピルビン酸トランスフェラーゼ (GPT)	アラニンアミノトランスフェラーゼ (ALT)
くん煙	燻煙
コンジローム	コンジローマ
さや	莢
ざんごう	塹壕
ざん壕	塹壕
ジスキネジー	ジスキネジア
ジストニー	ジストニア
しゃ血	瀉血
しゅう酸	シュウ酸
じゆく状	粥状
じょくそう	褥瘡
じんま疹	蕁麻疹
し緩	弛緩
し骨	篩骨
し眠	嗜眠
すう勢	趨勢
ぜい弱	脆弱
せき	咳
せん	癬
せんもう	せん妄
そう	叢

そう病	躁病
そう痒	搔痒
そしゃく	咀嚼
そ径部	鼠径部
たばこ	タバコ
たん	痰
たんぱく質	タンパク質
つる	蔓
でん部	殿部
でん粉	デンプン
てん絡	纏絡
とう骨	橈骨
どもる	吃る
のう	囊
ばい煙	煤煙
はさみ	鋏
パシー	パチー
ハンチントン舞蹈病	ハンチントン病
ひとみ	瞳
ひょうせつ	剽窃
ひ骨	腓骨
ひ素	ヒ素
び漫性	びまん性
ふるい	篩
ふ化	孵化
べんち	胼胝
ポリプ	ポリープ
まつげ	睫毛
ミオトニー	ミオトニア
よう素	ヨード
らい病	ハンセン病
ラセン	らせん
り患	罹患
るい瘦	るいそう
レセプター	受容体
れんさ	連鎖
レンサ	連鎖
阿片	アヘン
暗きよ	暗渠
煙草	タバコ
横痃	よこね
牡蠣	カキ
可そ性	可塑性
鈎	鈎
覚せい	覚醒
間歇	間欠
顔ぼう	顔貌
砧骨	きぬた骨

交叉	交差 (交叉)
骨粗しょう	骨粗鬆
骨髄	骨格
産じょく	産褥
子かん	子癩
屍体	死体
漆	ウルシ
遮へい	遮蔽
充てん	充填
上皮小体	副甲状腺
色盲	色覚異常
食慾	食欲
新生物	腫瘍
震顫	振戦
靱帯	靱帯
洗滌	洗淨
繊維	線維
絨毛	線毛
鼠蹊部	鼠径部
蛋白質	タンパク質
痴呆	認知症
蜘蛛膜	くも膜
蔦	ツタ
殿粉	デンプン
澱粉	デンプン
尿路性器系	腎尿路生殖器系
函数	関数
非定型抗酸菌	非結核性抗酸菌
副睪丸	精巣上体
米ぬか	米糠
偏頭痛	片頭痛
弁逆流	弁閉鎖不全
縫帯	包帯
蜂窩織炎	蜂巣炎
沃素	ヨウ素
沃素	ヨード
螺旋	らせん
硫黄	イオウ
湾曲	彎曲
佝僂病	くる病
佝僂病	くる病
喘ぎ	あえぎ
嵌入	陥入
彎	弯
攪	攪
鬱血	うっ血
鬱病	うつ病
瀰漫性	びまん性

畸形	奇形
疝痛	仙痛
疣	いぼ
癩菌	らい菌
癩癧	てんかん
辜丸	精巣
砒素	ヒ素
糜爛	びらん
羸瘦	るいそう
羸?	るいそう
腔	腔
臀部	殿部
莓	莓
蓆酸	シュウ酸
蔗糖	ショ糖
蛔虫	回虫
訥る	吃る
頸	頸
鼠	鼠
鼠径部	鼠径部
齲齒	う齒
齲食	う食
精神薄弱	知的障害
精神薄弱	精神遅滞
精神病院	精神科病院
精神分裂病	統合失調症
_囊	そ囊
_氣	あい氣
_	啣
_	囊
_	囊
_	搔
_瘡	ぞ瘡
_痒	搔痒
_	瘦
_疽	ひょう疽
_	せつ
_	軀
_	齟
癰	よう
譫妄	せん妄
日支	肢
臍ぞう	臍臓
鬱血	うっ血
うつ血	うっ血
間塞	閉塞
皮フ	皮膚
懷死	壊死

委縮	萎縮
褥創	褥瘡
心源性	心原性
心厚性	心原性
脳硬塞	脳梗塞
消化管	消化管
消火管	消化管
壊疽	壊疽
姓	性
のがん	癌
悪性新生物	癌
イ頸	首吊り自殺
いっ頸	首吊り自殺
縊頸	首吊り自殺
縊首	首吊り自殺
縊死	首吊り自殺
?型	2型
?型	1型
StanfordA	大動脈解離StanfordA
大動脈乖離	大動脈解離
A型大動脈解離	大動脈解離StanfordA
ナルコース	ナルコース
ラルコース	ナルコース
憎悪	増悪
アルツハイマー認知症	アルツハイマー型認知症
くも膜出血	くも膜下出血
くま膜下出血	くも膜下出血
クモ膜下出血	くも膜下出血
みぎ	
ひだり	
右	
左	
低酸素脳症	低酸素性脳症
うっ血性心不全	うっ血性心不全
うっ血心不全	うっ血性心不全
うっけつ性心不全	うっ血性心不全
ㇿ血性心不全	うっ血性心不全
えん下性肺炎	嚥下性肺炎
がん悪液質	癌性悪液質
がん性	癌性
クロイツフェルトヤコブ病	クロイツフェルトヤコブ病
ごえんせい肺炎	誤嚥性肺炎
肺炎	肺炎
(誤嚥性)肺炎	誤嚥性肺炎
びまん性大細胞型B細胞リンパ腫	びまん性大細胞型B細胞性リンパ腫
びまん性大細胞B細胞性リンパ腫	びまん性大細胞型B細胞性リンパ腫
びまん性大細胞性B細胞性リンパ腫	びまん性大細胞型B細胞性リンパ腫
びまん性大細胞性B細胞リンパ腫	びまん性大細胞型B細胞性リンパ腫

びまん性大細胞柄B細胞リンパ腫	びまん性大細胞型B細胞性リンパ腫
一酸化炭素中毒死	一酸化炭素中毒
急性一酸化炭素中毒死	一酸化炭素中毒
急性一酸化炭素中毒	一酸化炭素中毒
COPD急性増悪	慢性呼吸不全急性増悪
胃(\D+)癌	胃癌
悪期	悪性
下顎歯肉(\D+)癌	下顎歯肉癌
両側肺炎	肺炎
両肺炎	肺炎
線癌	腺癌
(\D+)非浸潤性乳管癌	非浸潤性乳管癌
([^非]+)浸潤性乳管癌	浸潤性乳管癌
患性	悪性
第動脈	大動脈
解離性大動脈流	解離性大動脈瘤
拡張性心筋症	拡張型心筋症
幹細胞癌	肝細胞癌
感染症心内膜炎	感染性心内膜炎
胆肝癌	胆管癌
胆管	胆管
管内胆管癌	肝内胆管癌
肝間部胆管癌	肝門部胆管癌
肝質性肺炎	間質性肺炎
肝門部(\s+)胆管癌	肝門部胆管癌
肝門部(\D+)胆管癌	肝門部胆管癌
間接性肺炎	間質性肺炎
悪質液	悪液質
癌悪液質	癌性悪液質
癌悪液質	癌性悪液質
腹膜園	腹膜炎
急性くも膜下出血	くも膜下出血
急性一酸化炭素中毒	一酸化炭素中毒
急性呼吸(\D+)不全	急性呼吸不全
白血症	白血病
氏不全	心不全
急性室不全	急性心不全
新不全	心不全
筋萎縮性	筋萎縮性
筋萎縮性側管硬化症	筋萎縮性側索硬化症
筋萎縮性側策硬化症	筋萎縮性側索硬化症
筋萎縮側索硬化症	筋萎縮性側索硬化症
加仏	K
高K血圧	高K血症
骨髓異形性症候群	骨髓異形成症候群
骨髓異型性症候群	骨髓異形成症候群
骨髓異型成症候群	骨髓異形成症候群
骨髓異形症候群	骨髓異形成症候群

骨髓異形性症候群	骨髓異形成症候群
骨髓異常形成症候群	骨髓異形成症候群
骨髓性異形成症候群	骨髓異形成症候群
腺維	線維
腎盂	腎盂
上向結腸癌	上行結腸癌
腸管膜	腸間膜
心原性脳梗塞症	心原性脳梗塞
心原性脳梗塞症	心原性脳梗塞
心原性脳梗塞症	心原性脳梗塞
心原性脳梗塞症	心原性脳梗塞
多系統萎縮症	多系統萎縮症
胆管Ca	胆管癌
Na血症	ナトリウム血症
Ab血症	アルブミン血症
低体温症	低体温
特発性肺腺癌	特発性肺線維症
脳卒中	脳卒中
播種	播種
播種性血管内凝固症	播種性血管内凝固
播種性血管内凝固異常	播種性血管内凝固
播種性血管内凝固異常	播種性血管内凝固
播種性血管内凝固異常	播種性血管内凝固
播種性血管内凝固症候群	播種性血管内凝固
廃用性症候群	廃用症候群
癱用症候群	廃用症候群
敗血症	敗血症
敗血症	敗血症
敗血症	敗血症
敗血症	敗血症
非閉塞性腸間膜虚血	非閉塞性腸間膜虚血
非閉塞性腸管虚血	非閉塞性腸管虚血
非閉塞性腸管虚血	非閉塞性腸管虚血
腹膜Ca	腹膜癌
慢性腎不全	慢性腎不全
老すい	老衰
老スイ	老衰
嚥下傷害	嚥下障害
攝食障害	摂食障害
濾胞性悪性リンパ腫	濾胞性リンパ腫
膠芽腫	膠芽腫

参考資料4-2 未コード化傷病名変換ルール

不詳	R99
不明	R99
縊死	T71
急性心臓死	I461
慢性心不全急性増悪	I509
低酸素脳症	G931
不詳の内因死	R99
誤嚥	T179
致死性不整脈	I498
虚血性心不全	I255
左乳癌	C509
右乳癌	C509
なし	R99
縊頸	T71
クモ膜下出血	I609
慢性心不全の急性増悪	I509
右肺癌	C349
慢性心不全増悪	I509
大動脈解離	I710
虚血性心疾患（推定）	I259
脳幹出血	I613
虚血性心疾患の疑い	I259
心臓死の疑い	I461
不詳の病死	R99
縊頸	T71
左肺癌	C349
蘇生後脳症	G931
重症肺炎	J189
心疾患の疑い	I519
急性心筋梗塞の疑い	I219