別添資料３

実データを対象としたIrisによる原死因確定実験

概要

突合死亡票 DB の一部を対象に、死亡個票の全ての病名に ICD-10 コードが機械付与可能であったもの 3,267 件を対象として、Iris によって原死因を確定し、実際の確定原死因との比較実験を行った。Iris では一切の付帯情報について考慮していないが、結果として約 9 割（全体に対して 88.4%、付帯情報が無いものについては 92%）の死亡個票に対し、Iris による原死因確定結果が実際の確定原死因と一致した。またさらなる精度向上のために必要な処理も明らかになり、我が国のオートコーディングツールを模すものとして十分に代替可能であることが判明した。

はじめに

本研究全体では、原死因確定プロセスにおける人手チェックの効率化のための機械学習の適用を目的としている。我が国ではオートコーディングツールによって確定された仮原死因に対し、死亡票に何らかの付帯情報があった場合、人手チェックにより最終的な原死因の確定が行われている。ここで機械学習の援用により、原死因の変更の有無を予め高精度に予測することができれば、人手チェックを大幅に効率化することができる。

しかし、そのためには、WHO が定める原死因確定ルールに従い、仮の原死因を確定する必要がある。現状では厚生労働省内で実際に用いられているオートコーディングツールを用いることはできないため、この挙動を模すツールとして、世界的に利用されているオートコーディングツールである Iris をその代替として用いることを想定している。

これまで、「疾病、傷害および死因統計分類提要　ＩＣＤ－10 準拠　第２巻　インストラクションマニュアル　<総論>」（以下、インストラクションマニュアル）中に記載されている原死因コーディング例（Ⅰ欄Ⅱ欄中の病名+選択するべき原死因の傷病名）を対象に Iris を用いた原死因確定実験を行った。その結果から、Iris の原死因コーディングの精度が約 8 割であることが示された。しかしながらインストラクションマニュアルの特性（マニュアルとして載せるべき判断の難しい例がまとめられている）から、実データを用いた場合さらに精度が上がる可能性があることも示唆された。

そこで、本実験では、実データ（実際の死亡票・死亡個票データ）を用い、Iris による原死因確定の挙動を調査することを目的とする。Iris は日本語に対応していないため、入力として自由入力病名を ICD-10 コード化したものを用いる必要がある。これには【別添資料

1】にて示した通り、標準病名マスターとの単純な文字列一致で行えるものから、簡単な言語処理を施すもの、人手による高度な判断を必要とするものと複数のレベルが存在している。

インストラクションマニュアル中の事例を対象とした場合、そのレベルの違いにより Step0 処理（ICD-10 コード付与可能 21.8%）、Step1 処理（ICD-10 コード付与可能 48.7%)、Step2 処理(ICD-10 コード付与可能 100%）と ICD-10 コードが付与できる件数が増加するが、いずれも、Iris にて正しく原死因が付与できる割合は 約 8 割であった。インストラクションマニュアルでは、難しい事例が多く掲載されているが、実データではより簡単な事例の頻度が多いと考えられる。実際、実データでは【別添資料 2】にあるように Step0 処理だけで ICD-10 コード付与可能なものが 21.8%から 43.6% へ大幅に増加している。このことから、Iris にて原死因が確定できる割合もインストラクションマニュアル事例（約 8 割）から増加すると考えられ、本実験ではこれについて調査を行う。

実データ

本実験でいう実データとは、平成 27 年～平成 30 年の死亡票とオンライン報告された死亡個票の調査票情報を結合した突合死亡票 DB を指す。その中でも、Step0 の簡単な処理だけで死亡票内の全病名に ICD-10 コードが付与できたもの 3,267 件を対象とした。

実験

実データを用いて Iris で原死因コーディングを行う。

方法

手順1.　Access データの生成

Iris は Access データを読込んで原死因コーディングを行う。そのため、突合死亡票 DB 中から以下の情報を抽出し、Access データを生成する。

- 対象者情報テーブル
  - ID
  - 生年月日
  - 死亡年月日
  - 性別
- 病名テーブル
  - 欄番号（Ⅰ欄ア：0, イ：1, ウ：2, エ：3, Ⅱ欄：5）
  - ICD-10 コード

手順2.　Iris による原死因コーディング

実験で用いる Iris のバージョンは Iris Version 5.6.0-Y2019S1 であり、原死因コーディング部分を行う MUSE のバージョンは MUSE 2.7 である。

手順 1.で生成した Access データをもとに Iris のバッチ（一括）処理機能を利用し原死因コーディングを行う。ただし、以下の ICD-10 コードに関しては Iris が対応していないため変換を行う。

- 2 型糖尿病（2 件）：E11→ E119
- 糖尿病（15 件）：E14 → E149
- 下肢閉塞性動脈硬化症（1 件）：I7020 →I702
- 急性呼吸不全（40 件）：J9609 →J960
- 慢性呼吸不全（8 件）：J9619 →J961
- 2 型呼吸不全（2 件）：J9691 →J969
- 呼吸不全（43 件）：J9699 →J969

結果

3,267 件中、

（ア）原死因が付与されたデータ　：3,250 件（約 99.5%）

（イ）原死因が付与されなかったデータ　：17 件（約 0.5%）

死亡票に記載される原死因（以下、確定原死因）を正解データとしたとき、Iris による原死因との比較を行った結果、（ア）に対しては、

・正解　：2,888 件（　/3,250 ≒ 88.9%）

・不正解　：362 件（　/3,250 ≒ 11.1%）

全 3,267 件に対しては、

・正解　：127 件（　/3,267 ≒ 88.4%）

・不正解　：29 件（　/3,267 ≒ 11.6%）

であった。また、（イ）に関して、

・”Rejected code”と記載されたもの　：2 件（　/3,267 ≒ 0.06%）

・全病名が空欄（対象者情報のみ）であったもの：15 件（　/3,267 ≒ 0.46%）

・付帯情報欄に記載がある：10 件

・海外での死亡　：7 件

・死体検案書の添付の旨が記載されている　：2 件

・その他記載がある　：1 件

・付帯情報欄に記載がない（確定原死因は全て R99）：5 件

であった。

考察

実データに Iris を適用した際の精度は約 9 割であることがわかった。

原死因が付与されなかったデータに関して、”Rejected code”と記載された 2 件については、多系統萎縮症（G903）の ICD-10 コードが Iris には存在していないことが原因として考えられる。また、病名が空欄であった 15 件は現時点で Iris での解決は困難であると考えられる。

不正解のデータから、考察したことは以下の通りである.

i. 確定原死因にアルファベットの 5 桁目が存在（43/ 362 件）

該当する ICD-10 コードを表 1 に示す。Iris と確定原死因を比較した際に、確定原死因の末尾にアルファベットが記載されているものが存在した。これは、日本特有の記載である可能性が高く、対処が必要である。例えば、確定原死因のアルファベットを削除する（情報の粒度を荒くする）と、22 件が正解に転じた。

ii. Ⅰ Ⅱ欄と確定原死因の ICD-10 コードの桁数が異なる（40/ 362 件）

該当する ICD-10 コードを表 2 に示す。ほとんどの場合、Iris による ICD-10 コードの方が多い。これに関しては桁数の調整により解決できると考えられ、全件正解にすることが可能であると考えられる。

表 1　確定原死因にアルファベットの 5 桁目が存在

確定原死因	病名	件数	確定原死因	病名	件数
G122A	筋萎縮性側索硬化症	7	S065A	急性硬膜下血腫 慢性硬膜下血腫	6 1(I620)
J100B	インフルエンザ A 型	1(J101)	S066A	外傷性クモ膜下出血 くも膜下出血	1 1(I609)
J152A	MRSA 肺炎	2	S269A	心臓損傷	1
J841B	特発性間質性肺炎 特発性肺線維症	6 2	S273A	肺挫傷	1
J841C	肺線維症	2	S328A	骨盤骨折	1
K803B	総胆管結石性胆管炎 急性閉塞性可能性胆管炎	1 1(K830)	S720A	大腿骨頚部骨折	1
K805B	総胆管結石	1	T142A	骨折	1
S062A	脳挫傷	6			

表 2　ⅠⅡ欄と確定原死因の ICD-10 コードの桁数が異なる

確定原死因	病名	ⅠⅡ欄(IRIS)	病名	件数
L89		L899	仙骨部褥瘡	1
I48		I489	心房細動	18
I48		I482	慢性心房細動	1
I48		I480	発作性心房細動	2
E460	栄養失調	E46		1
E461	低栄養失調	E46		1
C80		C809	脊索腫	1
C80		C800	原発不明癌	10
A09		A099	出血性大腸炎･急性腸炎	1･1
A09		A090	感染性腸炎･腸管感染症	1･2

iii. Iris の主傷病名欄に ICD-10 コードがふられている（69/ 362 件）

Iris は、原死因の他に主傷病を出力する。そして、そこの欄に記載される ICD-10 コードが原死因である場合が存在しており、まとめると以下のようになった。

・完全一致：39 件（/3250＝1.2%）

・アルファベットの 5 桁目除外で一致：18 件（/3250＝0.6%）

・不一致：12 件（/3250＝0.4%）

したがって、主傷病名欄の ICD-10 コードにも着目する必要があると考えられる。

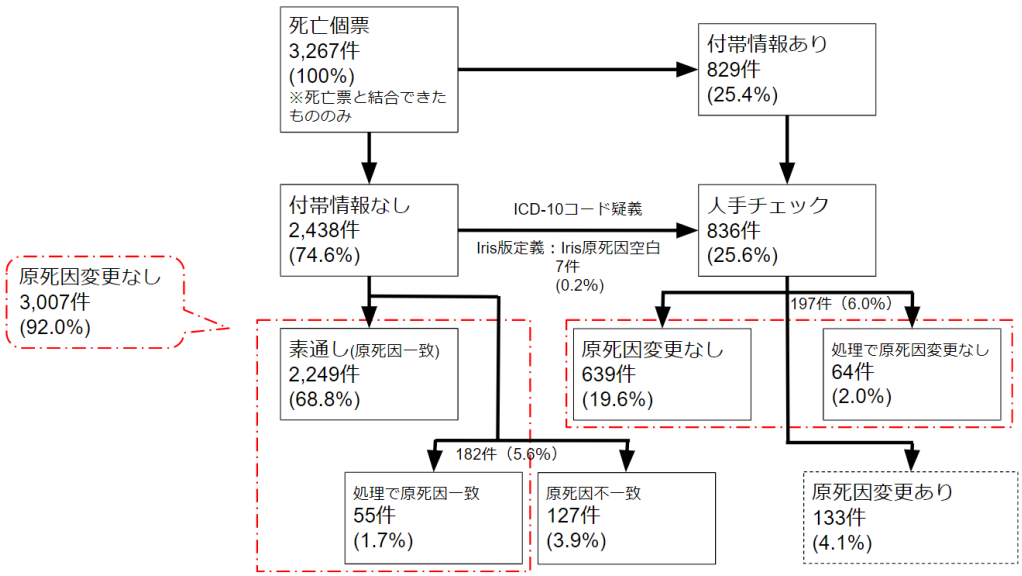

図　1　実データの構造の概要（推定）

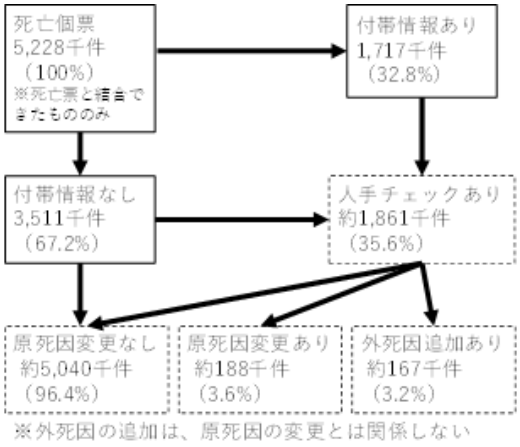

図　2　死亡票の流れ図

以上より、不正解の原因 i ～ iii に関しては、以下のような処理により正解が増加すると考えられる。今回の場合は、

1.アルファベットの消去　　：22 件

2.桁数の調整　　：40 件

3.Iris の主傷病名欄の ICD-10 コードを採用：57 件

であり、119 件（/3,250=3.7%）正解が増加し、総正解数は 3,007 件（/3,250=92.5%）になった。残りの不正解のデータに関しては、Iris との原死因コーディング方法の違いも挙げられるが、付帯情報の影響も考えられる。

3,267 件のデータから実データの構造の概要を推定し、図 1 を作成した。四角中には、実データの状態、件数、（全体における割合）の順で記載がされており、厚生労働省のヒ

アリング等から作成した死亡票の流れ図（図２）を参考にしている。ここで、実線の四角は確定情報を、等間隔の点線の四角は推定情報を、赤の点線の四角は原死因変更なしに分類されるものである。

図１より、死亡票と結合できた死亡個票データつまり突合死亡票 DB から抽出された 3,267 件（100%）の実データの中で、付帯情報がないものは 2,438 件（74.6%）、付帯情報があるものは 829 件（25.4%）であった。付帯情報がないものに関して、Iris で原死因コーディングを行った結果、確定原死因と一致したものは 2,249 件（68.8%）であり、一致しなかったものは 182 件（5.6%）であった。確定原死因と一致しなかったもののうち処理を行うことによって確定原死因と一致したものは 55 件（1.7%）であり、一致しなかったものは 127 件（3.9%）であった。一方、付帯情報がないもののうち原死因が付与できなかったものは 7 件（0.2%）であり、これを ICD-10 コード疑義として付帯情報があるものと足し合わせ、人手でチェックするものとしたのは 836 件（25.6%）である。人手でチェックするもののうち、確定原死因と一致したものつまり原死因に変更がないものは 639 件（19.6%）であり、一致しなかったものつまり原死因に変更があると推察されるものは 197 件（6.0%）であった。確定原死因と一致しなかったもののう処理を行うことによって確定原死因と一致したものは 64 件（2.0%）であり、一致しなかったものつまり原死因に変更があると推察されるものは 133 件（4.1%）であった。図中より、処理をせずに原死因が一致するものは 2,888 件（88.4%）であるが、処理を行うことによって原死因が一致するものも原死因変更なしとするとその数は 3,007 件（92.0%）であった。

以上の結果と図２を比較すると、付帯情報がある割合が少ないことがわかる。これは、対象となる 3,267 件が予備実験により得られているためであると考えられ、予備実験による ICD-10 コード付与の精度（現在は 43.6%）が高まるにつれて増加すると考えられる。また、図１での推定値はある１か月のデータをもとに作成しているため、月による数の変動も考えられる。しかし、人口動態死因オートコーディングシステムが利用できない状況下で、Iris を用いた場合でも、人口動態死因オートコーディングシステムとの大きな乖離がないと考えられる。したがって、より多くの実データを用いた検証も必要であるが、今後方針を大きく変えることなく「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」の機械学習パートにおける、付帯情報がある死亡票に対して人手でチェックすべきかしなくてもよいかを二値に分けるための、学習データが確保できると考えられる。

まとめ

3,267 件の実データを用いて、原死因コーディングを行った。その結果、約 9 割の精度で原死因が確定できることが分かった。また、付帯情報の有無による実データの概要も把握することができた。人口動態死因オートコーディングシステムが利用できない状況下でも、代替として Iris を用いた原死因確定プロセスの検証は可能であると考えられる。今後は、対象の実データを増やし、図 1 の付帯情報有りのデータを図 2 の割合（32.8%）まで収集できるようにしていく予定である。