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））
総括研究報告書（令和２年度）

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

研究代表者　今井　健　（東京大学大学院医学系研究科 准教授）

研究要旨

人口動態調査は国勢調査と並ぶ国の主要統計で公衆衛生施策の中心的資料である。本研究は原死因確定に関する調査を行い、我が国での原死因データ収集における課題を抽出し、ICD-11 における死亡診断書や死亡統計ルールの動向を調査すると共に、原死因確定作業に対する機械学習の適用可能性について調査・検討を行うことを目的とする。本年度は、死亡票の実データを対象に、文字列処理と自動 ICD-10 コード付与を行った上で、オートコーディングツール IRIS を適用し、約 65%の死亡票に対し、仮原死因を確定した。またこの仮原死因の変更の有無、外因や母側病態のコード追加の有無の割合について明らかにすると共に、機械学習用データセットを作成した。さらに、「何らかの付帯情報に影響され、仮原死因が変更に至るか否か」を予測する2値分類モデルによる機械学習の結果、基礎的な情報しか用いないベースライン手法でも、非常に高い精度(Accuracy90%) で変更の有無を判別可能と判明した。本手法を発展させることで、これまで人手確認によって行われてきた原死因確定作業の大幅な効率化、負荷軽減が図れると期待される。

研究分担者

香川璃奈

筑波大学医学医療系 講師

研究協力者

大井川仁美

奈良県立医科大学 MBT 学講座博士課程

大江和彦

東京大学大学院医学系研究科 教授

今村知明

奈良県立医科大学公衆衛生学講座 教授

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。今後 ICD-11 を国内適用するにあたっては原死因データを適切に収集・分析し、国際比較可能なデータを提供することが求められている。レセプト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。また我が国では高齢化が進み死亡者数の増加が見込まれることから、より正確で効率の高いデータ収集の方法の検討が求められている。

そこで、本研究は、原死因確定に関する調査を行い、我が国での原死因データ収集における課題を抽出し、ICD-11 における死亡診断書や死亡統計ルールの動向を調査すると共に、原死因確定作業に対する機械学習の適用可能性について調査・検討を行うことを目的とする。

B．研究方法

B-1）原死因確定プロセスにおける課題の抽出

すでに昨年までで明らかになっている原死因確定プロセスの詳細割合について、平成27年～平成 30 年の死亡票・死亡個票実データの分析を追加し、アップデートを行った。

B-2）機械学習の適用可能性調査

死亡票・死亡個票実データ(平成 27～30 年、約 520 万件)に対し、各種の文字列処理と標準病名マスターを利用して自動 ICD-10 コーディングを行うシステムを開発した。

また、全病名に ICD-10 コードが振られたものについては、IRIS に入力し、仮原死因コードを決定すると共に、確定原死因コードと比較を行った。その際に、IRIS と国内のコーディングルールの差異を吸収する処理も行った。以上の一連の処理は自動化し、Docker 並びに仮想マシンにおいて実行可能なシステムとして実装した。

また IRIS 処理の結果、何らかの修正処理が必要なケース、不要なケースに分類し、その内容について集計を行うと共に、機械学習のための教師データの作成を行った。

最後に、このデータを用い、何らかの付帯情報が存在するケース 50 万件を対象とし、付帯情報によって影響を受けて「IRIS が付与した仮原死因」が変更されるか否か、を２値分類する機械学習を行った。本年度はまずはベースライン手法とし、① Ⅰ欄 II 欄各病名の ICD10 コード、② 付帯情報の各項目の有無、③ IRIS が付与した仮原死因、を入力データとし、分類器学習モデルとして汎用的な勾配ブースティング決定木の一種である XGBoost を用いて、仮原死因が変更されるか否かを予測するモデルを構築した。

B-3）ICD-11 における死亡診断書や死因統計ルールの動向調査

我が国の現行の死亡統計では ICD-10 を元にした WHO による原死因選択ルールが適用されている。しかし 2018 年 6 月に WHO が ICD-11 をリリースした今、ICD-11 における死亡統計の動向は今後の我が国への ICD-11 適用に際し重要である。本年度は昨年度に引き続き WHO 並びに日本 WHO-FIC 協力センターの関係者へのヒアリングによってこの動向調査を行った。

尚、本研究では倫理面への配慮は必要としない。

C．研究結果

C-1）原死因確定プロセスのアップデート

昨年度までに判明している原死因確定プロセスについて、本年度は追加で提供を受けた平成３０年度のデータを加え、総数約 517 万件に対して改めて原死因確定の流れを計算し直し、アップデートを行った。結果を図１に示す。昨年度までと大きな変更はないが、(1) 何らかの付帯情報があるもの (32.4%) の内訳はオートコーディングツールによるコーディングエラーあり (11.5%)、なし(20.9%) となっており、(2) 付帯情報がない 67.6% のものから、コーディングエラーのある 3.2%を加えた 35.6%について目視確認を行っていると改めて推計された。これまでのヒアリング結果とも齟齬は無い。

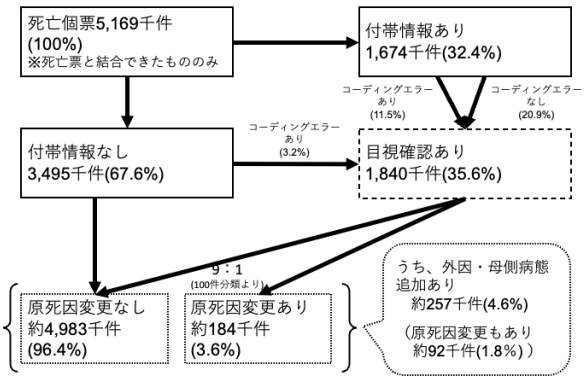

図１：原死因確定プロセス（改訂版）

詳細については、「別添資料１ 原死因確定プロセス調査(昨年度調査のアップデート)」を参照されたい。

C-2) 機械学習の適用可能性

本年度は、平成 27～30 年の 4 年間の死亡票・死亡個票突合データ（約 520 万件）に対し、全Ｉ欄 II 欄病名に対する自動 ICD コーディングを行い、IRIS による仮原死因確定処理を行った。さらに、何らかの付帯情報が存在するケースを対象とし、付帯情報によって影響を受けて「IRIS が付与した仮原死因」が変更されるか否か、を２値分類する機械学習を行った。

(1) まず、4 年間の死亡票・死亡個票突合データについては、「届出地と事件簿番号」の組み合わせを各死亡案件のキーとして用いたが、複数回存在するものが 29,410 件存在したため、これを除いた約 517 万件のものを用いた (C-1,図１も同様)。この複数存在するキーはヒアリングの結果、早期提出に起因するものとのことであった。

(2) 次に、全ての病名に対し、読点の削除、複数病名列挙の展開、文字の正規化などの文字列処理を施した上で標準病名マスターを用いて ICD-10 コーディングを行った。現状では、74%の病名に対して自動 ICD-10 コーディングが可能で、全ての病名が ICD-10 コード化された死亡票は 65%であった。この「65%」という割合は十分とは言えず、死亡票の特徴による偏りを可能な限り排除するために、この割合をなるべく向上させる必要がある。そこで、本年度は (A) この ICD-10 コーディング可能割合を向上させる処理、(B) ICD-10 コードが付与されたとして、その後 IRIS による仮原死因確定、変更の有無予測の機械学習、と進む処理、とを分け、並行して研究を行った。

(A)については、研究分担者香川璃奈が中心に担当し、様々なルールベースの前処理を施すことで、最終的に「全ての病名記載に対し ICD-10 コーディング可能である死亡票の割合」は 80%にまで増加した。詳細については、香川璃奈分担研究報告書「**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**」を参照されたい。

一方、これと並行し、(B) ICD10 コードが付与されて以降の処理を行ったため、以下は本年度当初の 65%の死亡票 (全ての記載病名が ICD10 コーディングできたもの) を対象とした結果について記す。

(3) IRIS による仮原死因確定処理は全件に対して行うと約 10 日間程度かかるため、その後の調査は、約 10%をランダムサンプリングした 50 万件に対して行った。50 万件中、全病名が ICD-10 コーディングされ、IRIS への入力として利用できたもの は約 32 万件(65%)である。

また、IRIS は ICD-10 の国際版に準拠しており、日本国内で適用されている独自コードは実装されていない。このような事例については IRIS が出力する仮原死因コード、死亡票における確定原死因コードのそれぞれについて修正処理を行った。

表１に、この調査対象とした 320,112 件 (全 50 万件の調査対象セットのうち、全病名に ICD-10 コードが付与でき、IRIS に入力できたもの) の内訳を示す。IRIS が付与した仮原死因と、死亡票の確定原死因とを比較した結果、変更があるものは全体の約 10%、「何らかの付帯情報がある」81,725 件に限定すると、この中の約１5% (11,987 件)に

対し原死因の変更を行うべきであることが判明した。IRIS の処理時間は約２０時間程度であった。

仮原死因	付帯情報あり	付帯情報なし	計
変更あり	11,987 (3.7 %)	17,337 (5.4 %)	29,324 (9.2 %)
変更なし	69,738 (21.8 %)	221,050 (69.1 %)	290,788 (90.8 %)

表１：調査対象セット中の仮原死因変更の割合

詳細は、**別添資料２「Iris による仮原死因付与処理」**を参照されたい。

また、原死因の変更以外にも、人手によるチェックで修正が行われるものがある。これが損傷や外因の影響（ICD-10 第１９章）に対する「外因コード」(V01-Y98)の追加、周産期における母側病態コードの追加などの「コード追加」である。つまり原死因には変更がなくてもこのような補足コードが追加されることがあり得る。これについても、人手によるチェックが行われていることから、AI システムによる支援の対象と捉え、細分化した結果が以下の表２である。

仮原死因	コード追加必要	コード追加不要
変更あり	1.9 %	12.8 %
変更なし	4.3 %	81.0 %

表２：付帯情報が存在するものに対するコード変更・追加の割合

表２は「何らかの付帯情報」が存在するもの(つまり必ず人手のチェックに回るもの)のうち、仮原死因の変更の有無、何らかのコード追加(外因, 母側病態)の発生の有無の割合を示している。これによると、仮原死因の変更もなく、コード追加の必要もないのは約８割であり、残りの２割は何らかの修正処理が必要であることになる。

つまり、

1) 最初にこの２割と８割 (1:4) の２値分類を行い、「対処の必要がない８割」をまず削減する
2) 次に２割のものについて、仮原死因の変更の有無、コード追加の必要性の有無を判別し、可能であればその変更先や追加コード自体の推薦を行う、

という**２段構えの支援システムが必要であることが判明した**。また、これらの結果は同時に、今後支援システムに必要な分類器の学習に必要な教師データとして用いられるものである。

(4) 次に、以上の処理で得られた学習用データを元に、何らかの付帯情報が存在するケース 50 万件を対象とし、付帯情報によって影響を受けて「IRIS が付与した仮原死因」が変更されるか否か、を２値分類する機械学習を行った。本年度はまずはベースライン手法とし、① Ⅰ欄 II 欄各病名の ICD10 コード、② 付帯情報の各項目の有無、③ IRIS が付与した仮原死因、を入力データとし、分類器学習モデルとして汎用的な勾配ブースティング決定木の一種である XGBoost を用いて、仮原死因が変更されるか否かを予測するモデルを構築した。学習に使用された特徴(Feature)は 1577 次元であり、出力は「仮原死因の変更あり・なし」の２次元である。今後の精度比較のための最も単純なベースライン手法であるため、付帯情報については内容を考慮せず、単に各項目に記載があるかないかだけ(0/1)を用いている。

学習の結果、**Accuracy90.3%で仮原死因の変更の有無が予測できることが判明した**。結果を表３（四分表）ならびに表４（各種指標）に示す。

		正解		
	変更	なし	あり	
予測	なし	13545	1161	14706
	あり	423	1209	1632
		13968	2370	16338
	accuracy:	90.3%		

表３：仮原死因変更の有無予測結果（4 分表）

	なし	あり
recall	97.0%	51.0%
precision	92.1%	74.1%
F	0.94	0.60
specitivity	51.0%	97.0%

表４：仮原死因変更の有無予測結果（各種指標）

また、分類に寄与した重要な特徴量について下記図２に示す。結果として年齢、備考欄の記載の有無、I501 コードの存在、その他付言すべき事柄の記載の有無、手術年月日の有無、性別、という順になった。I501, J188, A415 といったコードは、I 欄 II 欄病名にこれらのコードになる病名が存在することを意味し、付帯情報については内容を加味していないため、単にどの項目に記載があるかを意味している。

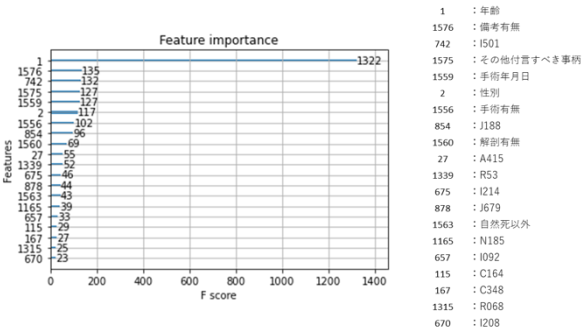

図２：分類に寄与した重要な特徴量

詳細については「**別添資料３機械学習を用いた原死因変更有無判定**」を参照されたい。また本機械学習手法による変更有無予測結果と正誤については、一部を本統括・分担報告書全体の末尾にある「**別添資料**」中に示したので合わせて参照されたい。

C-3) ICD-11 における死亡診断書や死亡統計ルールの動向

本年度も昨年度に引き続き関係者へのヒアリングの結果、現段階では WHO は ICD-11 における死因統計ルールについて公表しておらず、また Iris の ICD-11 対応も作業が開始されているもののリリースまでは当分時間がかかるということであった。原死因選択のルールについては基本的な考え方は踏襲されるものと思われる。しかし ICD-10 に比べて大幅に粒度が細かい疾病分類体系となった ICD-11 では Iris における原死因選択ルールテーブルが大幅に変更になり、これに合わせ我が国でのこれまでのオートコーディングシステムでのルールベースも大幅な変更を余儀なくされると予想される。次年度以降引き続き動向を注視することが必要である。

D．考察

本年度の成果で、死亡票実データの約 65%を対象として、IRIS による仮原死因確定処理を行い、確定原死因と比較することで、原死因コードの変更の割合、また外因や母側病態コードの追加割合が明らかになった。またこの結果により今後の分類器学習に必要な教師データセットが得られ、本年度の目標を達成した。

何らかの対処が必要なもの（原死因コードの変更、外因や母側病態コードの追加）は両者を合わせて 2 割であり、最初に 1:4 の分類タスクにより「対処の必要がない 8 割」を除去した後に、細かな対処内容の分類を行う 2 段階処理が適していると考えられた。**最初の分類タスクが高精度に行えるだけでまず 8 割の人手処理を削減できることが**

期待でき、また 2 段階目の処理によって残りの 2 割に対する人手作業の効率化が図れると期待できる。

本年度行った仮原死因の変更の有無に関する２値分類学習では Accuracy90%と非常に高い精度での判別が可能であった。ベースライン手法であるため、各付帯情報については「記載の有無」しか用いておらず、その内容については一切考慮していない。その段階でこのような高い精度が実現できたことは驚くべきことであり、今後、付帯情報の内容を TF・IDF や BERT で学習された内部表現ベクトルなどを用いてモデルに組み込むことで一層の精度向上が見込まれ、非常に有望な手法と考えられた。特に、変更や追加の有無だけでなく、変更後・追加コードの提案まで行う場合は手術や解剖所見の文章、備考欄や付言欄の自由記述文章を対象とした BERT による汎用言語モデルの獲得と利用が必要と考えられ、現在実験中である。既にベースラインの手法で高精度を実現してはいるが、前述の 2 段階処理の第 1 段階目の分類精度としてはまだ向上の余地があり、次年度の課題である。

また並行して行った自由入力病名の ICD10 コーディング(分担研究)では 80%の死亡票につき全記載病名がコーディング可能と大幅に能力が向上しており、次年度ではこのデータを元に IRIS 処理、ならびに機械学習処理を行う予定である。

E．結論

本年度研究では、死亡票の実データに対して IRIS を適用し、約 65%に対し仮原死因を決定した上で今後の分類器学習のための教師データが得られた。また原死因コードの変更、コードの追加割合についても明らかにした。付帯情報の影響による仮原死因変更の有無についての2値分類ではベースライン手法にて Accuracy90%での判別が可能と判明した。

F．健康危険情報

なし

G．研究発表

1. 明神 大也, 大井川 仁美, 香川 璃奈, 今村 知明, 今井 健. 死因統計の精度と効率性の向上に向けた我が国の原死因確定課題の抽出. 医療情報学 40(Suppl.):674-676, 2020.
2. 大井川 仁美, 明神 大也, 香川 璃奈, 今村 知明, 今井 健. 原死因確定プロセスにおける IRIS の国内導入可能性に関する基礎的な検討. 医療情報学 40(Suppl.):677-682, 2020.
3. 今井 健, 明神大也, 大井川仁美, 香川璃奈, 今村知明. 原死因確定作業についての実態・問題点の把握、ならびに正確・効率性向上に向けた機械学習の適用可能性と課題に関する調査研究. 厚生の指標, 2020;67(3): 17-24, 2020.

H．知的財産権の出願・登録状況

なし