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（令和元年度）

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、ヒアリングと死亡票の実データを元にした集計によって、原死因確定プロセスにおける課題と処理の流れの概要を明らかにした。また、オートコーディングツール Iris について調査を行い、原死因選択ツールとして利用可能であることを明らかにすると共に、現在の人手作業の大半を占める「付帯情報による原死因コード変更確認」の機械学習による支援に向け準備を整えた。

研究分担者

明神大也

奈良県立医科大学病理診断学講座 医員

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。

原死因データ収集における課題については、既に申請者らは平成 30 年度厚生労働統計協会調査研究委託事業において、ヒアリン

グ調査・関連資料分析などを通じ基礎調査を行ってきている。本研究ではこれを発展的に展開し、人口動態調査の実データを用いた統計解析と機械学習の適用を行うことで、より詳細な分析を行う。また、申請者らはこれまで厚労科研での ICD-11 改定の動向研究(H26-28, H29-31)に参画し、国内 ICD 検討会での活動を通じて ICD 改定の種々の問題を抽出・議論・意見発信を行ってきた。本研究では引き続きこれらの体制・国内関連学会と連携し、ICD-11 における死亡診断書・死亡統計ルールの動向調査を通じ、ICD-11 に準拠した原死因データ収集の際に必要な対応について知見を集積する。

本年度は初年度として、統計法 33 条に基づく人口動態調査情報の提供を受け、欧米諸国を中心に利用が進むコーディングツール IRIS の国内利用可能性について分析を行うと共に ICD-11 における死亡統計ルールについて動向調査を行った。

B．研究方法

本研究は(1) 原死因確定プロセスの調査と課題の抽出、(2) 機械学習の適用可能性調査、(3) ICD-11 動向調査、(4) セキュリティのあり方検討、の 4 つの柱から構成されている。これらは分断した分担制ではなく、それぞれに研究代表者、研究分担者、研究協力者が参画する、という体制をとっている。研究初年度である令和元年度ではこのうち (1)～(3) に研究課題が存在しており、以下順に述べる。

B-1) 原死因確定プロセスにおける課題の抽出

原死因データ収集における課題については、既に申請者らは平成 30 年度厚生労働統計協会調査研究委託事業において、ヒアリング調査・関連資料分析などを通じ基礎調査を行ってきている。本研究ではこれを発展させ、より詳細な分析を行うため、厚生労働省関係者へのヒアリングと共に、統計法第 33 条に基づく目的外利用申請によって平成 27 年～平成 30 年の死亡票・死亡個票データの提供を受け、実データを元にした分析を行った。また統計法 22 条によりこのうち一部について、厚生労働省にて人手チェックに回った調査票情報の提供を受け、このサンプリング結果に対する分析と合わせることで、原死因確定の流れを明らかにした。

B-2) 機械学習の適用可能性調査

厚生労働省のオートコーディングシステムでは、I 欄 II 欄傷病名に ICD-10 コードを付与し、その組み合わせから WHO が定める原死因の選択ルールに則って、ルールベースで仮の原死因が決定される。その後、コード付与の疑義や何らかの付帯情報(手術解剖の有無と所見、外因死の追加事項、生後 1 年未満死の追加事項、その他付言すべきことがら等)が存在する場合には職員による人手確認を経て、最終的な原死因コードが決定される。このシステムを用いて仮の原死因コードが導出できれば、付帯情報が最終的な原死因コードの決定(仮の原死因コードからの変更の有無)へ与える影響が明らかとなり、この部分を機械学習で支援することが可能となる。しかし残念ながら上記オートコーディングシステムは公開されておらず、これを利用することができない。そこで本年度は、無料で公開され、広く欧米約 20 カ国で用いられているオートコーディングシステムである Iris をその代替として用いることができるか調査を行った。

Iris は日本語をサポートしていないが、I 欄・Ⅱ欄病名を ICD-10 コーディングした結果を入力すれば、原死因選択ルール処理の動作を確認することは可能である。

A) そこで、実データの提供を受ける前の事前実験として、疾病、傷害及び死因の統計分類提要 ICD-10(2013 年版)準拠

第二巻 総論 （インストラクションマニュアル） 中に詳細に記述されている原死因選択事例集 （全 156 件） を対象に、各傷病名記述を人手で ICD-10 コーディングし、これを Iris への入力としてインストラクションマニュアル中の正解と比較した。

B) 次に、死亡票・死亡個票の実データ提供を受けた後は、この実データの一部を対象として、I 欄・II 欄病名に対する ICD-10 コーディングを行った。これを Iris への入力とし、確定原死因との比較を行った。

B-3) ICD-11 における死亡診断書や死亡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 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における課題

まず、関係者へのヒアリングを通じ、原死因確定プロセスにおける課題として、大きく次の 2 つが挙げられた。

(1) オートコーディングシステムで原死因がルールベースで決定できない事例

- 死亡票の I・II 欄傷病名が自由入力であるため、辞書マッチングで ICD-10 コードが付与できないことがある
- 原死因選択ルールに合致しない、疑義がある （「老衰」が年齢と合っていない、希少疾患である等）

(2) 何らかの付帯情報が存在する場合

- 付帯情報が存在する場合は必ず人手での確認処理に回され、必要があればオートコーディングシステムが出力した仮の原死因コードを修正している。
- これには以下のような多様な事例が存在している。
  - ✧ I 欄(ア)に「肺炎」、手術欄に「胃悪性腫瘍切除術・1 週間前」とある。
    → 本来 I 欄(イ) に「胃癌」と書くべきとみなしてこれを選択する。
  - ✧ I 欄(ア)で「損傷」、手段・状況欄で、自殺,飛び降り、あるいは交通事故とわかるとそれを優先 (外因等)。
  - ✧ 「細菌性肺炎」とあるが解剖欄の情報で菌の種類が分かる場合詳細化する。
  - ✧ 「飛び降り自殺」とあるが、I 欄内で産後うつによる影響であると分かると妊産婦死亡のフラグを付与する。

次に、目的外利用申請によって提供を受けた死亡票・死亡個票の実データを結合した。以下これを「突合死亡票データベース(DB)」と呼ぶ。死亡票と死亡個票が結合できたのは平成 27,28,29,30 年の順に、125 万件、128 万件、132 万件、137 万件、であり、この総数 522 万件のうち、何らかの付帯情報があるものは 171 万 (32.8%)、付帯情報なし 351 万 (67.2%)であった。

同じく統計法 22 条に基づき提供を受けた人手チェックに回った調査票情報によると、1 ヶ月間の死亡票約 11 万のうち、人手チェックに回ったものは 35.6%であった。先の「付帯情報があるもの」の割合 (32.8%) と合わせて考えると、「付帯情報なし」のうちの約 3%は「疑義がある」(上記(1) のケース)と推定され、「付帯情報あり」(上記(2)のケース)、と合わせた約 36% が人手の確認に回っていることが判明し

た。これは月の件数として約4万件程度である。一方、残りの約 64%はオートコーディングシステムが決定した原死因コードがそのまま確定される。

従って、付帯情報あり（(2) のケース）が人手確認の大半を占めており、これを計算機支援する必要があることが判明した。

また、人手チェックに回った調査票情報からのランダムサンプリング 100 件を分析した結果、「付帯情報あり」が 80 件、「疑義がある」が 45 件、両方に該当するものは 24 件であった。またこれらを仮原死因と確定原死因が異なった（人手チェックで変更となった）ものは 10 件、外因に関わる符号や母側に関わる符号が追加されたものは 9 件であった。サンプリング数が少ないものの、人手チェックの結果原死因が変更される割合は少なく、チェックに回ったもののうちの約 1/10、全体のおおよそ 3～4%程度であると推定された。

付帯情報やコーディングエラーの確認の結果原死因が変更される割合は少なく、大多数は変更ないと考えられることから、原死因の変更の有無を高精度に予測する機械学習アルゴリズムの導入が人手チェック作業効率化のために有効と考えられた。

詳細については、**分担報告書「死亡に関わる調査票情報提供に基づいた原死因確定プロセスにおける課題の抽出」**（明神大也）を合わせて参照されたい。

C-2）機械学習の適用可能性

C-1）での分析により、「何らかの付帯情報があるケース」を計算機支援することが必要である。この際、人手確認を行っても付帯情報が特に影響を及ぼさなかった場合は、オートコーディングシステムが決定した仮の原死因コードがそのまま確定される。従って、この仮の原死因コードが必要となるが、死亡票データにはオートコーディングシステムの出力結果が含まれておらず、また当該システムは公開されていない。

そこで今回、代替として Iris が利用できるか調査を行った。

(A) インストラクションマニュアル事例に対する調査結果

まず、死亡票・死亡個票の提供を受ける前の事前実験として、我が国の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選択事例を対象に、傷病名の ICD-10 コードを入力とし Iris で原死因選択を行った。

結果、全 156 中、正解 127 件（81.4%）、不正解 29 件（18.6%）となった。

不正解であったものの内訳は

- 悪性新生物に関する死因記載の解釈に問題がある（16/30 件）
- 及び と 又は の区別の不備（8/30 件）
  （Iris は and / or を区別せずどちらにしても and として解釈を行ってしまう。）
- その他（6/30 件）

となっていた。いずれも主に「元の傷病名表現を正確に ICD-10 コード化できなかった」ことに起因する問題である。正解率は 80%であるが、インストラクションマニュアル中の事例は判断に迷うような複雑・イレギュラーなケースを載せていることが多く、死亡票の実データなど一般的なケースへ適用する際には十分厚生労働省のオートコーディングツールの代替として利用可能と考えられた。

一方、Iris への入力としては ICD-10 コードが必要であるが、標準病名マスターを用いても、自由記載の病名に対し ICD-10 コーディングを完全に自動で行うことは難しい。そこで、ICD-10 コーディングを行う際に、原文記載まま(Step0)、文字列処理を要するもの（Step1)、人手判断を要するもの （Step2） の 3 段階の基準を定め、分類を行った。インストラクションマニュアルの事例では、Step1 までで約 2 割、Step2 までの処理で約 5 割のものに ICD-10 コードを

付与することができ、Step0～2 の処理結果いずれに対しても、Iris にて約 8 割以上の精度で原死因が確定可能であった。

この詳細については【別添資料１】「インストラクションマニュアル事例を対象とした Iris による原死因確定実験」を参照されたい。

(B) 死因病名に対する ICD-10 コーディング実験結果

上記の結果、実データ（突合死亡票 DB）を解析するにあたり、病名に対する ICD-10 コード付与が課題となることが判明した。そこで、実データ中の自由記載病名を対象に、ICD-10 コード付与可能な割合の調査を行った。

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。

この詳細については、【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」を参照されたい。

(C) 実データを対象とした Iris による原死因確定実験結果

上記(B)の結果、Step1)までの処理により死亡個票の全ての病名に ICD-10 コードが機械的に付与可能であったものの一部(3,267 件)を対象とし、Iris にて原死因を確定し、実際の確定原死因との比較を行った。先の(A)はインストラクションマニュアル中の事例を対象としたもので、難しい事例が多いが、実データでは容易なケースが多いと予想される。

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。

この詳細については、【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」を参照されたい。

以上の結果をまとめると、標準病名マスターと簡単な文字列処理を組み合わせることで 65%程度までは全病名に対し ICD-10 コードが付与でき、Iris への入力にかけることができると判明した。また Iris も我が国のオートコーディングツールの代替として十分に利用可能であることが判明したため、これを利用すれば「付帯情報に影響されて仮原死因が変更されるか否か」を機械学習により自動分類するための学習用セットを十分に確保できそうである、という感触を得た。

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、実データを元にした原死因確定プロセスの処理の流れをその割合の概略が明らかとなり、人手で確認されているものが約 36%程度であること、その大半は「何らかの付帯情報があるもの（約 32.8%)」であることが判明し、この部分を計算機支援する必要があると考えられた。その際には WHO の原死因選択ルールに基づいて、オートコーディングシステムが決定する「仮の原死因コード」が必要であるが、傷病名コードを ICD-10 コード化して入力することで Iris が利用可能であることが判明した。

Iris が正しく原死因決定できなかったものは、元の傷病名の表現を正確に ICD-10 コード変換できないことが主な原因であった。従って「付帯情報の影響」に特化して機械学習するためには、「全て傷病名が ICD-10 コード付与できる事例」のみに絞って学習を行えば良い。これによって付帯情報を考えない場合での原死因選択結果（仮の原死因コード）を得ることができる。具体的には標準病名マスターの利用によって全ての I 欄 II 欄傷病名に ICD-10 コードが完全付与できるケースのみに絞ることが考えられる。

本年度の成果によって、次年度以降の機械学習適用のための基盤が整備された。今後はまず「何らかの付帯情報があり」、かつ「全ての傷病名が ICD-10 コード付与可能なもの」を対象にし、これを Iris に入力して得られた仮の原死因コードと、死亡個票の確定原死因コードとの比較を行う予定である。この一致/不一致を教師データとして用いることで、付帯情報による原死因コード変更への影響を学習することが可能と考えられる。仮に原死因コードが変更される場合、「付帯情報によって何のコードに変更されるか」、まで予測出来るモデルを構築することが理想であるが、仮に「変更されるか否か」の２値分類を高精度に分類するモデルでも実用上の効用は大きい。なぜならば「変更されない」と高精度に予測された事例は人手で確認する必要が無くなるため、人手作業対象が大幅に削減されると考えられるからである(本年度の結果からは約 1/10 程度しか変更が発生せず、残りは確認する必要が無い)。次年度以降は、まずこの２値分類モデル、さらには変更後のコード予測モデル、へと開発を進めて行く予定である。

E．結論

本年度研究では、ヒアリングと死亡票の実データを元にした集計によって、原死因確定プロセスにおける課題と処理の流れの概要を明らかにした。また、Iris の利用可能性について調査を行い、付帯情報による影響の機械学習に向け準備を整えた。

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