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（令和元年度）

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

研究分担者　明神 大也　（奈良県立医科大学医学部附属病院・医員）

研究要旨

日本では益々人口減少・高齢化の進展が予想されるが、今後の ICD-11 導入に際し死因統計の正確性の担保・より一層の効率性向上を図るためには、現行の原死因確定プロセスの課題を抽出することが重要である。これまでのヒアリング調査で、オートコーディングツールを利用した現行の原死因確定ロジックでは、死亡票の約4割に対し目視確認が必要であることが判明しているが、実データに基づく詳細調査は未だ十分に行われていない。

本研究では死亡票・死亡個票の実データを用いて、現行の死因確定プロセスにおける課題点についての調査を行った。具体的には、人口動態調査票情報の提供を受け、平成 27 年～30 年の死亡票・死亡個票を対象とし、結合処理を行って基礎的な統計量を得た。また、オートコーディングツール処理の結果、目視確認が必要とされた死亡票の一部を抽出し、目視確認の原因や原死因の変更について調査を行い、原死因確定への影響について推計を行った。

死亡票と死亡個票の結合処理により得られた約 520 万件のうち、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）が含まれていた割合は約3割であった。また目視確認に回る理由は付帯情報がある場合以外にも、死因を ICD-10 コードに変換できない、死因に病名以外の外因が含まれている、などの場合が存在していた。目視確認の対象となる死亡票は全体の約 3 分の 1 程度で、そのうち原死因が変更になる割合は約 10 分の1程度であると推定された。

本研究では、実データを用いて現行の原死因確定ロジックの課題を明らかにした。ICD-11 導入にあわせ、より高精度で効率性の高いロジックの開発が必要であるが、このためには自然言語処理による病名正規化処理と、原死因変更の有無を高精度に予測する機械学習アルゴリズムが有効と考えられた。今後はこの開発を進めていく予定である。

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。

本研究班（研究代表者・研究分担者・研究協力者）チームによるこれまでの先行研究として、平成３０年度 厚生労働統計協会 調査研究委託事業「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」において、死亡個票はオートコーディングツールにかけられて原死因が割り振られ、その一部は人手チェックに回り、必要に応じて原死因が変更され、確定原死因になることが分かっている。また、オートコーディングツールを利用した現行の原死因確定ロジックでは、死亡票の約４割に対し目視確認が必要であることが判明しているが、実データに基づく詳細調査は未だ十分に行われていない。

そこで本分担研究では原死因確定プロセスにおける課題の抽出を行い、(1)原死因確定調査および(2)人口動態調査情報の取得と統計処理を行った。

Ｂ．研究方法

人口動態調査情報の取得と統計処理

【別添資料４】に示した集計方法で集計を行うことを企画し、統計法 33 条に基づき、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の利用申請を行った。調査票情報の申請時点で以下の集計方法を考えた。

① 死亡個票では、付帯情報の有無を確認した後、「死亡の原因」に含まれるⅠ欄・II 欄病名の全てについて標準病名マスターおよびこれを用いた自然言語処理により ICD-10 コードを付与できたものを対象とする。
② 死亡票はそのまま読み込むこととする。
③ そして①の ICD-10 コードを用い､欧州で利用されている原死因確定ツール IRIS を用いて導出した原死因と、②の死亡票に含まれる確定原死因コードを比較し、合致したものをデータセット１、合致しなかったものをデータセット２とする。
①の死亡個票と②の死亡票を結合する条件として、死亡個票は処理年月・届出地（都道府県・保健所・支所符号・市区町村（種類）・市区町村（順位））・事件簿番号、死亡票は調査年・提出年月・届出地・事件簿番号とする。
④ その後、開発したアルゴリズムをテストセットデータに適用し、原死因を導出する。その結果と死亡票の確定原死因・外因符号の結果を照らし合わせ、精度検証と課題確認、またアルゴリズムの改善を繰り返し行う。

そして調査票情報の提供を受けて基礎的調査を行った。

なお、【別添資料５】に示す通り、死亡票では、調査年、提出年月、届出地、事件簿番号、性別、出生年月日時分、死亡年月日時分、原死因、外因符号、母側病態、単多胎別、妊娠週数、母の生年月日、前回の妊娠、子の数の項目の利用申請を行った。また、【別添資料６】に示す通り、死亡個票は、死亡届出地の都道府県と市町村、事件簿番号、処理年月、備考欄記述有無、死亡した人の都道府県と市町村、死亡したところの種別、「死亡の原因」に含まれるⅠ欄・II 欄の原因と期間、手術の部位及び所見、手術年月日、解剖の部位及び所見、死因の種類、障害が発生した年月日時分、傷害が発生したところの種別、傷害が発生したところ、その他の記述、傷害発生場所、手段及び状況、生後 1 年未満での病死の病態・異状の詳細、その他付言すべき事柄、備考欄の項目の利用申請を行った。

さらに本研究では、統計法 22 条に基づき、人手チェックに回ったリストの調査票情報の提供を受けた。死亡票はオートコーディングツールにかけられて仮原死因が確定する。その際の死亡票に付帯情報が入っておらず、且つコーディングエラーが発生しなければ仮原死因が確定原死因になる。一方、付帯情報が入っているまたはコーディングエラーが発生していれば人手チェックに回る。コーディングエラーは何らかの理由で死因に ICD-10 コードが振られなかった場合、または死因に外死因（病死以外）が含まれる場合に発生する。そこでオートコーディングツール実行後の死亡票のうち、ランダムサンプリングされた 100 件の人手チェックに回ったリストの調査票情報の提供を受け、基礎的調査を行った。

倫理面への配慮

本研究では個人情報や動物愛護に関わる調査・実験は行わない。ただし研究の遂行に当たっては、各種法令や「人を対象とする医学系研

究に関する倫理指針」を含めた各種倫理指針等の遵守に努めた。

C．研究結果

死亡票のレコード数は平成 27 年、28 年、29 年、30 年の順に、1,301,379 件、1,319,030 件、1,351,944 件、1,374,469 件（合計 5,346,822 件）であった。死亡個票のレコード数は平成 27 年、28 年、29 年、30 年の順に、1,249,469 件、1,280,853 件、1,325,955 件、1,374,780 件（合計 5,231,057 件）であった。これらを SQL サーバに格納【別添資料７】し、死亡票と死亡個票を結合【別添資料８】した。結合条件は処理年および届出地の都道府県・保健所・市区町村・事件簿の番号とした。

その結果、死亡票と死亡個票が結合できたのは平成 27 年、28 年、29 年、30 年の順に、1,248,057 件（男性 645,711 件/女性 602,346 件）、1,280,808 件（男性 661,501 件/女性 619,307 件）、1,325,413 件（男性 683,577 件/女性 641,836 件）、1,373,589 件（男性 705,560 件/女性 668,029 件）であった。合計 5,227,867 件（男性 2,696,349 件、女性 2,531,518 件）であった。

このうち、［死亡の原因］－［I 欄］－［ア欄］の原因・期間に入力があったのはそれぞれ 5,219,044 件（99.8%）・5,143,096 件（98.4%）であった。［イ欄］の原因・期間に入力があったのはそれぞれ 1,861,677 件（35.6%）・1,445,585 件（27.7%）、［ウ欄］の原因・期間に入力があったのはそれぞれ 393,799 件（7.5%）・312,152 件（6.0%）、［エ欄］の原因･期間に入力があったのはそれぞれ 73,542 件（1.4%）・54,670 件（1.0%）、［II 欄］の原因・期間に入力があったのはそれぞれ 1,744,402 件（33.4%）・1,628,459 件（31.1%）であった。

死亡の原因以外（以下「付帯情報」という。）に記載があったのは 1,716,961 件（32.8%）であった。

原因・期間ともに自由入力であったため、多数の類似表記が確認された。例えば原因が誤嚥性肺炎の場合、誤嚥性肺炎（90.4%）、嚥下性肺炎（4.4%）、誤えん性肺炎（1.7%）、誤燕性肺炎（1.2%）、反復性誤嚥性肺炎（0.3%）、肺炎（誤嚥性）（0.2%）などが 40 種類以上の表記が存在していた。期間についても１カ月の場合、１ヶ月、約１ヶ月、１カ月、約４週間、間もなく１ヶ月などが存在していた。

また、本研究では人手チェックに回ったリストの調査票情報の提供を受けた。1 カ月間の死亡票 約 113 千件のうち、人手チェックに回ったのは 40 千件で、割合は 35.6%であった。そのうちランダムサンプリングした 100 件のうち、付帯情報が含まれていたのが 80 件、コーディングエラーが含まれていたのが 45 件、付帯情報とコーディングエラーの両方が含まれていたのが 24 件であった。

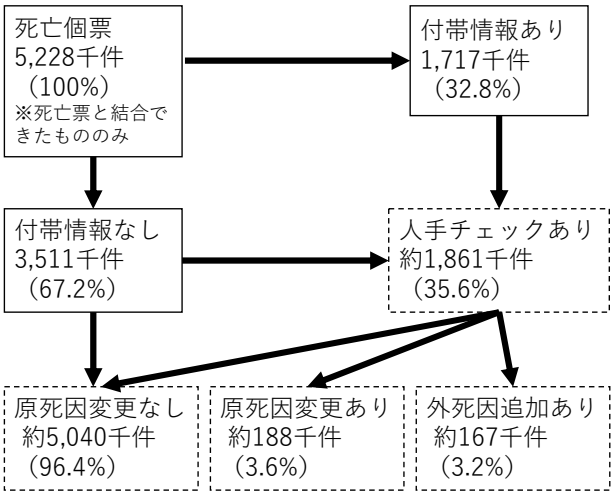
人手チェックに回った死亡票のうち、仮原死因と確定原死因が異なった（人手チェックで変更になった）のは 10 件、外因に関わる符号または母側に関わる符号が追加されたのは 9 件、変更なしが 81 件であった。

D．考察

死亡個票に比べて死亡票の件数が約 2%（4 年間で 116 千件）多かった。これは、死亡個票にはオンライン報告されたもののみ含まれるためと考えられた。

また、死亡票・死亡個票の調査票情報提供および人手チェックリストの調査票情報提供により、仮原死因を付与された死亡票は図１の流れになると考えられた。人手チェックに回る割合 35.6%は実数ではなく、1 ヶ月の抽出死亡票に対する値である。また、原死因の変更や外因の追加が行われる割合も、そこからランダム抽出された 100 件に対するものであり、統計的な信頼性は必ずしも高くはない。しかし、付帯情報があることにより人手チェックに回る事例（32.8%）以外にも、付帯情報は無いがコーディングエラーが生じる事例（約 3%）があり、合わせた結果として人手チェックが 35.6%に対し行われている、と考えられた。また、人手チェックの結果原死因が変更される割合は少なく、チェックされたものの約 1/10、全体のおおよそ 3～4%程度であると推定された。

なお、図１において外枠が実線の場合は死亡票・死亡個票から算出した実データで、点線の場合は抽出された人手チェックリストから推察した予測値を示す。また図１中で、外因の追加は原死因の変更とは関係しない。



（図１　原死因確定の流れ）

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。

一方、死亡の原因・期間は自由記述になっており、前述のように、多様な表現が存在している。今後 ICD-11 導入に合わせ、高い精度で効率性の高い原死因確定ロジックへ円滑に移行していく必要があるが、仮に国際的に広く用いられているオートコーディングツールである Iris を活用する場合には、この自由記述に対する自然言語処理による高精度な ICD コード付与が重要となると考えられた。これは、現行の我が国でのオートコーディングシステムで、人手チェックに回るコーディングエラーを減少させる上でも、重要である。

E．結論

本年度研究では、実データを用いて現行の原死因確定ロジックの流れ、また課題を明らかにした。ICD-11 導入にあわせ、より高い精度で効率性の高いロジックの開発が必要であるが、このためには自然言語処理による病名正規化処理と、原死因変更の有無を高精度に予測する機械学習アルゴリズムが有効と考えられた。今後はこの開発を進めていく予定である。

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