

令和3年政策科学総合研究事業(統計情報総合研究事業)
分担研究報告書

人口の健康・疾病構造の変化にともなう複合死因の分析手法の開発と
その妥当性の評価のための研究

死亡個票における死亡の原因欄の記載文字列の分析

研究分担者 篠原恵美子 東京大学大学院医学系研究科 特任助教

研究要旨 死亡個票データにおいて死亡の原因やその期間は自由入力データであり、統計処理に用いるためにはコード等への正規化が必要である。本年度は昨年度用いた正規化プログラムの変換精度の向上を図り、平成15年から令和2年までのデータの正規化を行った。また正規化処理の観点から死亡個票データの特徴を調査した。

A. 研究目的

死亡個票における死亡の原因欄は自由記載であるため、様々な表記ゆれが含まれており、例えば「虚血性心筋症」と「心筋虚血」のように表現が異なる場合や、「肺癌」と「左肺癌」のように側性の情報が付加される場合がある。これを統計処理するためにはコード化を行う必要がある。また、「肺癌、動脈硬化症」のように1つの欄に複数の病名が含まれる場合には、それぞれを別の病名として計数できなければならない。原因とペアで記録される期間も自由記載であり、正規化処理をしなければ統計処理ができない。しかし死亡調査票の数は年間100万件を超えており、全件を手で処理することは現実的ではない。そこで自然言語処理による自動正規化が有用と期待される。

昨年度は過去に開発した正規化プログラムを用い、平成15年から令和1年までの全データについて、死亡の原因欄に記載された内容のICD-10コード化および日数形式

への変換を行った。本年度は変換精度の向上を図ったうえで平成15年から令和2年までのデータの正規化を行った。また正規化処理の観点から死亡個票データの特徴を調査した。

B. 研究方法

(ア) 正規化プログラムの改修

変換プログラムではコード化に用いる知識として病名表記とICD-10コードの対応表を用いる。昨年度は病名コード表として標準病名マスター(以下、マスター)の当時最新版(Ver 5.05)を用いていた。しかし「認知症」のように現在では使用されなくなった病名はマスターの改訂時に削除されることがあり、コード化できない状態であった。本研究で扱う死亡個票データは古いものでは平成15年まで遡るため、現在では利用されなくなった病名もコード化したい。そこで、公開されている過去のマスターも全て統合して用いることとした。また、マスター

を補完する目的で、日本語の病名と ICD-10 コードの対応を含む万病辞書を導入した。

(イ) 死亡個票データの正規化

平成 15 年から令和 2 年までの全データ（オンライン分、17109860 件）に対し、アの改修を行った正規化プログラムを適用した。

(ウ) 死亡個票データの特徴調査

平成 15 年から令和 1 年までの全データについて、アの対策を行った正規化プログラムを適用し、記載内容を正規化の観点から分析した。

C. 研究結果

(ア) 正規化プログラムの改修

マスターの統合の結果、55735 の病名が得られ、最新のマスター（ver. 5.09）と比較して約 2000 の病名が追加されたものとなった。万病辞書からはこれに含まれない 90 の病名が追加された。

(イ) 死亡個票データの正規化

全ての年のデータについて、97%の死亡個票について I 欄アに少なくとも 1つの ICD-10 コードが付与された。

(ウ) 死亡個票データの特徴調査

原因欄フィールドのうち何らかの記載があるものは 3 割程度であり、そのほとんどが単一の病名のみを含んでいた。出現する病名の出典としてはマスターが 97%程度を占めていた。

D. 考察

昨年度の結果と比較し ICD-10 コードの付与割合が増えており、病名を追加したことによる効果と考えられた。死亡個票に記載される病名のほとんどはマスターに収載

されたものであり、コード化知識のリソースとしてマスターは適していると考えられる。ただし全てをカバーするわけではなく、精度の向上には追加の知識が必要であることが分かった。

E. 結論

昨年度用いた正規化プログラムの改修を行い、平成 15 年から令和 2 年までの死亡個票データについて正規化を実施した。

F. 健康危険情報

なし

G. 研究発表

1. 論文発表

なし

2. 学会発表

篠原 恵美子、別府 志海、林 玲子、石井 太. 「死亡個票における『死亡の原因』欄の記載文字列の分析」. 第 41 回医療情報学連合大会, 2021 年 11 月 20 日.

H. 知的財産権の出現・登録状況

なし