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））
分担研究報告書
ICD-11 の適用を通じて我が国の死因・疾病統計の向上を目指すための研究
「死亡個票における死亡の原因欄の記載文字列の自動正規化」

研究分担者　篠原恵美子 東京大学大学院医学系研究科

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。本年度はこれまでに開発してきた正規化プログラムを用い、平成 15 年から令和 4 年までのデータの正規化を行った。

A．研究目的

死亡個票における死亡の原因欄には病名が記載されるが、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。このような表記ゆれ関係にあるものを同一の病名として扱う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、全件を人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。

我々はこれまでに死亡の原因欄の変換プログラムを開発し複合死因の分析の前処理として用いてきた。本年度は同プログラムの変換精度の向上を図ったうえで平成 15 年から令和 4 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、および期間欄に記載された内容の日数形式への変換を行った。

B．研究方法

変換プログラムを、平成 15 年から令和 4 年までの全データ（オンライン分、20144189 件）に適用した。

コード化プログラムは病名と ICD-10 コードの対応表を用いるものであり、標準病名マスターの最新版を含め公開されているすべての版（ver. 2.10～5.13）を利用した。

（倫理面への配慮）

本研究では倫理面への配慮は必要としない。

C．研究結果

全ての年のデータについて、95%以上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。

D．考察

2000 万件を超えるデータについて、その 95%以上に自動で正規化を行うことが出来た。コードが付与されなかったデータについてはコード化知識を追加することで対応

可能であり、次年度の課題である。

Ｅ．結論

正規化プログラムの改修を行い、平成 15 年から令和 4 年までの死亡個票データについて正規化を実施した。

Ｇ．研究発表

１．論文発表

なし

２．学会発表

なし

Ｈ．知的財産権の出願・登録状況
（予定を含む。）

1. 特許取得

なし

2. 実用新案登録

なし

3.その他

なし