令和 2 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

死亡診断書における死亡の原因のコード化と期間表現の正規化

研究分担者　篠原恵美子　東京大学大学院医学系研究科　特任助教

研究要旨 死亡事故原票データにおいて死亡の原因やその期間についての情報は自由入力データであり、統計処理に用いるためには正規化が必要である。本年度は過去に開発した処理プログラムをベースにコード化知識の更新を行い、死亡の原因の ICD-10 コード・病名交換用コード化および期間表現の正規化を行った。

A. 研究目的

死亡調査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。

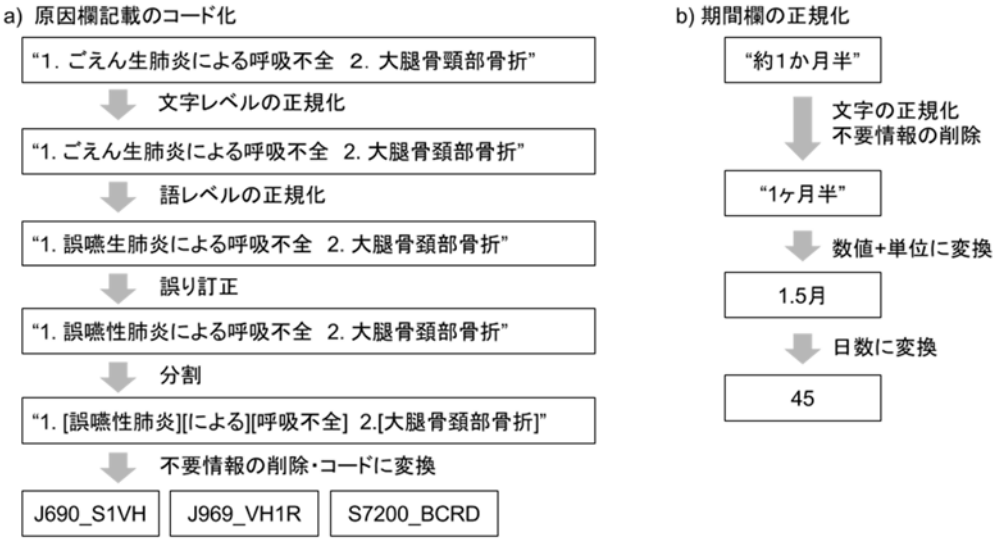
分担研究者は過去に死亡調査票の死亡の原因および原因欄に記載された内容の ICD-10 コード化および日数形式への正規化プログラムを開発した。今年度はこのプログラムで利用している病名コード表のアップデートを行ったうえで、平成 15 年から令和 1 年までの全データについてコード化・正規化を行った。

B. 研究方法

1. テキストデータの抽出

平成 15 年から令和 1 年の死亡票（調査票情報）15,725,292 件を対象とした。

死亡個票の電子ファイルは各項目が固定バイト長で格納された文字コード CP932 のテキストファイルであり、それよりも短いデータの場合には末尾が空白で埋められている。これを削除し、実際にテキストが含まれている部分のみを抽出した。また後の処理のため、文字コードを UTF-8 に変換した。抽出・変換したデータは全て SQLite3 のデータベースに 1 ファイル 1 テーブルとして格納した。ファイルごとにテーブルを分けたのは年ごとに列数が異なるためである。



2. 病名コード表のアップデート

原因欄と期間欄それぞれについて、自動でコード化・正規化を行うプログラムが実装しているアルゴリズムの概要を図 1 に示す。このうち、死亡の原因欄のコード化の「分割」ステップでは一般財団法人医療情報システム開発センターで公開されている ICD10 対応標準病名マスターを利用している。このマスターは年に数回更新されており、最新の Ver.5.05 をインストールした。

3. 原因欄のコード化

「死亡の原因 Ⅰ欄」(ア～エ)「死亡の原因Ⅱ欄」の「原因」を、前処理を行った上でマスターを利用して分割し、ICD-10 コードと病名交換用コードに変換した。前処理は文字の正規化(例. 頸→頚)・語の正規化(例. 十二腸→十二指腸)・誤り訂正(例. 十二腸→十二指腸)の 3 段階から成っている。データ量が非常に大きく、素朴な実装では実行時間が非常に長くなるため、形態素解析器の MeCab を利用し効率化を図った。

4. 期間欄の正規化

原因欄と同様に「死亡の原因 Ⅰ欄」(ア～エ)「死亡の原因Ⅱ欄」の「期間」に格納されているデータを対象とし、日数に変換した。元データとしてはこれらに加えて死亡日を利用した。死亡日は、年数などの時間幅を記録すべき「期間」に日付が記入されている場合があり、これを時間幅に変換するために必要である。

期間欄についても複数の情報が列記されていることがあるため、原因欄と同様に分解した。

図 1 コード化・正規化アルゴリズムの概要

5. 処理結果のまとめ

最後に、ICD-10 コード、病名交換用コード、期間(日数)を、年ごとに元となった死亡個票に列を追加する形でデータを出力した。

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欄に少なくとも 1 つの ICD-10 コードが付与された。

D. 考察

原因のコード化には病名マスターを利用しているが、カバーされない病名表記は独自の変換表を用意して対応している。このマスターと変換表でもまだカバーできていない範囲があり、今後改善が必要である。

また、期間正規化の結果がマイナスの数値になっているケースが見られた。この原因としては死亡個票の内容の誤りとプログラムの誤りの可能性があり、調査が必要である。前者に対してはそのようなケースをどう扱うかの検討が、後者に対してはプログラムの修正を行う予定である。

さらに、原因と期間の両方について、入力文字数に制限があるため、備考欄に続きが記載されているケースの存在が確認されている。このようなケースの割合の調査、対応する原因・期間と接続し再構成する方法の確立が必要である。

E. 結論

申請データ全数について死亡個票の原因欄および期間欄の基本的な正規化を自動で行った。