平成 27 年度厚生労働科学研究補助金（政策科学総合（統計情報総合）研究事業）
死亡個票統計における循環器疾患関連死因の妥当性に関する検討
（H27-統計-一般-006）分担研究報告書

# 死亡診断書における死亡の原因および期間表現の正規化

報告者（分担研究者）

篠原恵美子　　東京大学医学部附属病院 企画情報運営部 特任助教

**抄録**
死亡事故原票データが電子的な形態で利用可能となったが、死亡の原因やその期間についての情報は医師による自由記載であり、統計処理に用いるためには正規化が必要である。本年度は正規化のための基本的な処理の実装、および既存の病名集と自然言語処理技術を組み合わせた ICD-10 コードの自動付与を試みた。その結果、死亡個票のうち 89.7%に対し、ICD-10 コードを付与することができた。

## 【A．研究目的】

死亡票における死因は自由記載であるため、様々な表記ゆれが含まれている。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正規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。しかし死亡票の数は年間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を試みた。

## 【B．方法】

### （１）対象

2013年度の死亡個票（オンライン報告分 1,180,293件）における「死亡の原因　Ⅰ欄」「死亡の原因Ⅱ欄」の「原因」と「期間」に格納されているデータを対象とした。

### （２）方法

(a) テキストデータの抽出

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CP932形式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。

(b) 記載内容の正規化

原因欄と期間欄それぞれについて、自動で正規化を行うアルゴリズムを考案し、実装した。

(b-1) 「原因」の正規化

原因欄のテキストを正規化し、ICD-10コードを得る手法を実装した。

原因欄の記載は表記ゆれや複数の病名が含まれていることがあるため、まずこれを以下の方法で処理した。

- Unicode正規化（正規化形式KC）
- 一般的な医療分野の文字レベルの表記ゆれの解消（例. 頚→頸）
- 本データに頻出する表記ゆれの解消（例. 菅→管）

次に、医療情報システム開発センターから公開されているICD-10対応標準病名マスターの索引用語を用いて文字列を分割し、対応するICD-10コードの列に変換した。分割処理では形態素解析器のMeCabを用い、解析用辞書として上記マスターの索引用語のみを用いた。この結果から箇条書き番号および側性の情報を削除した。最後に、複数のICD-10コードが含まれる場合にはこれらを並列とみなし、分割した。正規化の結果はICD-10コードが付与された場合は当該ICD-10コードであり、その他の場合は以下の通りである。

- “EXT”外因死
- “UNK”「不詳」など
- “none”「なし」など
- “GG”その他空欄でないもの

(b-2)「期間」の正規化

期間欄も原因欄と同様に表記ゆれや複数の期間を含む場合があるため、漢数字からアラビア数字への変換、「約」などの語の削除を行った上で、正規化を行った。正規化の結果は、以下のいずれかである。

- 具体的な時間（例. 1.5月）
- +[単位]（例. 期間欄の記載が「数カ月」の場合「+月」）
- 不詳
- 長期間
- 短期間
- 短時間
- 日付（期間欄に日付のみが記載されている場合）

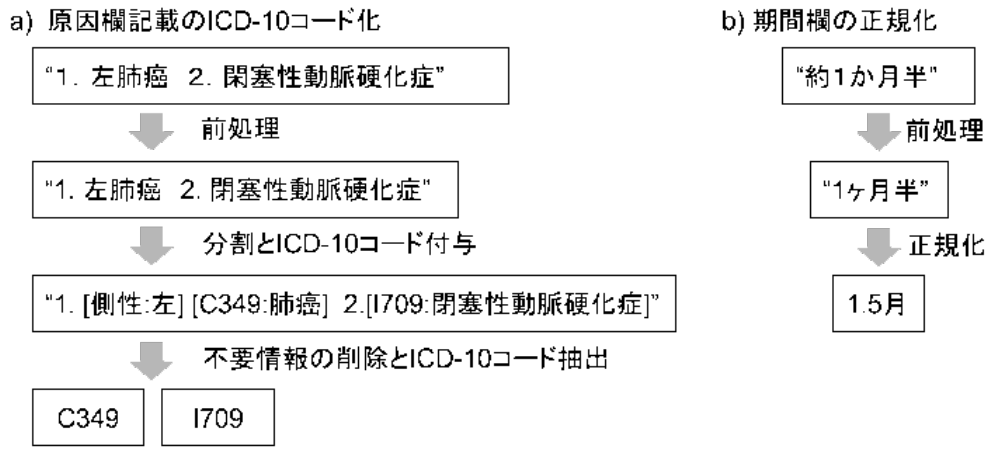


図 1. 正規化アルゴリズムの概要

【C. 結果】

実装したアルゴリズムを対象データに適用した結果を表 1 に示す。死亡個票のうち 89.7%（1,058,613 件）に対し、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】

大量の病名データを短期間で統計処理用に整形するために、本年度は文字レベルの正規化や並列表現の分解に注力して実装と自動コード化を行い、精度の向上を目指してアルゴリズムの修正を繰り返し行った。この過程で、個票データそのものにいくつかの問題点があることが伺われた。これは、個票データは表 2 の手順で作成されており、この各過程でエラーが発生しうるために起こるものである。本研究班が対象とするのは step 1 のエラーであるが、実際に用いるデータには step 2-6 におけるエラーが混入している。これらはそれぞれが異なる性質を持っていると考えらえるため、自動で完全な修正を行うのは困難であろう。このようなエラーを防ぐためには発生源入力、すなわち死亡診断書を作成する医師がコンピュータに入力を行い、第三者による転記を行わないというのが理想的な方法である。より簡便な方法としては死亡診断書を手書きではなく電子的に作成し印刷することが挙げられる。

また Step 1 に該当するエラーでは、原因欄に詳細な状況を記述していたり、期間欄に日付を記載していたりするなど、死亡診断書の記入マニュアルに従っていないケースが散見された。このようなエラーもコンピュータでの入力であれば比較的簡単に防止可能である。

【E. 結論】

死亡個票の原因欄および期間欄について、基本的な正規化を自動で行うことができるようになった。次年度は個々の病名の自動コーディング手法を検討し、より多くの個票を統計処理に用いることを目指す。

表 2. オンライン死亡個票の作成過程と起こりうるエラー

| step | 過程 | エラーの例 |
|---|---|---|
| 1 | 医師が記載内容を想起する | 「転倒」（状況を想起） |
| 2 | 医師が紙に記入する | 「血管」（書き間違い） |
| 3 | 保健所担当者が読み取る | 「三万パチー」（読み間違い） |
| 4 | 保健所担当者が記載内容を決める | 「記載なし」<br>（作業内容の不統一） |
| 5 | 保健所担当者がコンピュータでタイプする | 「万戦心不全」（タイプミス） |
| 6 | 保健所担当者がかな漢字変換の結果を確定する | 「配布善」（変換ミス） |

**表 1. ICD-10 コード付与結果**

✓は正規化結果に該当列が含まれることを示す。**例えば 2 行目は何らかの** ICD-10 **コード・**EXT**（外因死）、**GG**（未コード化）、**UNK**（不明）の 4 つを、Ⅰ欄に含む個票が** 15 **件、Ⅱ欄に含む個票が** 0 **件であることを表す。**

| ICD | EXT | GG | UNK | none | Ⅰ欄 | Ⅱ欄 |
|---|---|---|---|---|---|---|
| ✓ | ✓ | ✓ | ✓ | ✓ | 0 | 0 |
| ✓ | ✓ | ✓ | ✓ | | 15 | 0 |
| ✓ | ✓ | ✓ | | ✓ | 0 | 0 |
| ✓ | ✓ | ✓ | | | 726 | 0 |
| ✓ | ✓ | | ✓ | ✓ | 0 | 0 |
| ✓ | ✓ | | ✓ | | 310 | 0 |
| ✓ | ✓ | | | ✓ | 1 | 0 |
| ✓ | ✓ | | | | 6566 | 0 |
| ✓ | | ✓ | ✓ | ✓ | 3 | 0 |
| ✓ | | ✓ | ✓ | | 7002 | 0 |
| ✓ | | ✓ | | ✓ | 35 | 0 |
| ✓ | | ✓ | | | 85615 | 0 |
| ✓ | | | ✓ | ✓ | 17 | 0 |
| ✓ | | | ✓ | | 116030 | 0 |
| ✓ | | | | ✓ | 436 | 0 |
| ✓ | | | | | 841857 | 327061 |
| 小計 | | | | | 1058613 | 327061 |
| | ✓ | ✓ | ✓ | ✓ | 0 | 0 |
| | ✓ | ✓ | ✓ | | 39 | 0 |
| | ✓ | ✓ | | ✓ | 0 | 0 |
| | ✓ | ✓ | | | 844 | 0 |
| | ✓ | | ✓ | ✓ | 0 | 0 |
| | ✓ | | ✓ | | 398 | 0 |
| | ✓ | | | ✓ | 3 | 0 |
| | ✓ | | | | 14871 | 230 |
| | | ✓ | ✓ | ✓ | 0 | 0 |
| | | ✓ | ✓ | | 10909 | 0 |
| | | ✓ | | ✓ | 53 | 0 |
| | | ✓ | | | 78886 | 74965 |
| | | | ✓ | ✓ | 26 | 0 |
| | | | ✓ | | 13558 | 676 |
| | | | | ✓ | 1 | 2683 |
| | | | | | 2093 | 774679 |
| 小計 | | | | | 121681 | 853233 |
| 計 | | | | | 1180294 | 1180294 |