一般口演 | 知識工学
一般口演9
自然言語処理・テキストマイニング
2021年11月20日(土) 09:10 〜 11:10 E会場 (2号館2階222+223)

[3-E-1-07] 死亡個票における「死亡の原因」欄の記載文字列の分析

*篠原 恵美子[1]、別府 志海[2]、林 玲子[2]、石井 太[3] （1. 東京大学, 2. 国立社会保障・人口問題研究所, 3. 慶應義塾大学）
*Emiko Shinohara[1], Motomi Beppu[2], Reiko Hayashi[2], Futoshi Ishii[3] （1. 東京大学, 2. 国立社会保障・人口問題研究所, 3. 慶應義塾大学）

キーワード：Natural language processing, Vocabulary, death certificates

【背景】死亡診断書には死亡原因として、ア）直接死因、イ）アの原因、ウ）イの原因、エ）ウの原因、またこれらに影響を及ぼした傷病名等の5つの記入欄が用意されており、またそれぞれに対応する発症から死亡までの期間の記入欄がある。国の死亡統計ではこれらを元に決定した単独の原死因を用いるが、個票が研究利用可能となり、より詳細な分析が可能となった。その際、死亡原因の自由記載を病名コードに、期間を日数等に正規化する必要がある。
【目的】死亡診断書の悉皆データの原因・期間欄に対する自動正規化を実装して適用し、正規化処理の観点からこの自由記載データの特徴を明らかにする。
【方法】電子データとして利用可能な全データである平成15年から令和1年までの全死亡個票データを統計法に基づき申請し、材料とした。死亡原因の自動コーディングの実装として前処理および標準病名マスターから作成した辞書・万病辞書・追加の辞書による分割数最小法を、期間の自動正規化として有限状態機械を実装した。これを年ごとに材料に適用し、各辞書による病名のカバー率や期間表現を分析した。
【結果】17年分15725292件の個票データを用いた。各欄は必ずしも全てが記入されるわけではなく、「不詳」に該当する表現を削除すると65%前後が空欄であり、その割合は年とともに増える傾向にあった。実装した解析器を適用した結果、何らかの記入がある欄のうち9割前後から病名が抽出された。病名のうち、マスターに収載されているものが97%以上を占めていた。正規化に必要な情報は、原因欄・期間欄の文字列のみならず、死亡日・生年月日・備考欄が必要である場合があった。また、一つの欄に複数の原因・期間が記載されることがあるため、正規化処理中には原因欄と期間欄を同時に参照することが必要であった。

死亡個票における「死亡の原因」欄の記載文字列の分析

篠原 恵美子[*1]、別府 志海[*2]、林 玲子[*2]、石井 太[*3]
*1 東京大学、*2 国立社会保障・人口問題研究所、*3 慶應義塾大学

An analysis of “Cause of Death” fields written in death certificates

Emiko Shinohara[*1], Motomi Beppu [*2], Reiko Hayashi[*2], Futoshi Ishii[*3]
*1 The University of Tokyo, *2 National Institute of Population and Social Security Research, *3 Keio University

Background: In the death certificate, there are five fields for the cause of death: a) direct cause of death, b) cause of (a), c) cause of (b), d) cause of (c), and other significant conditions contributing to death. There is a corresponding "duration to death" for each column. While the national mortality statistics use the single underlying cause of death determined based on these data, more detailed analysis is possible based on the fields. In this case, it is necessary to normalize the free entry of the cause of death to a disease code and the period to a number of days, etc.

Purpose: To clarify the characteristics of the cause of death description from the perspective of normalization processing.

Method: The analysis target was all death certificate data from 2003 to 2019 in Japan, which are available as electronic data. Automatic normalization was developed and applied to this data and the results were analyzed. As the automatic coding of causes of death, we implemented a minimum number of partitions method using preprocessing, a dictionary created from the standard disease name master, Manbyo dictionary, and an additional dictionary. Finite state machine implemented as automatic duration normalization.

Result: We used total of 15725292 certificate data for 17 years. Around two-thirds of the fields were left blank, and one-thirds contained at least one disease name, almost of which were covered by the master. The information required for normalization was not only the text in the cause and duration fields, but also the date of death, date of birth, and remarks fields were sometimes necessary. In addition, since multiple causes or durations could be listed in one field, it was necessary to refer to the cause and duration fields simultaneously during the normalization process.

Keywords: Natural language processing, vocabulary, death certificates

1. 緒論

日本では人が死亡すると医師または歯科医師により死亡診断書が作成される。そこには死亡した人の名前等のほか、死亡の原因となった病名などが記載される。「死亡の原因」の記載欄はⅠとⅡの2つに分かれており、Ⅰはさらにア）直接死因、イ）アの原因、ウ）イの原因、エ）ウの原因、の 4 つに分かれている。Ⅱは死因には直接関係しないがこれらの経過に影響を及ぼした傷病名等を記載する欄である。またそれぞれに対応する「発症から死亡までの期間」の欄がある。この死亡診断書は紙媒体として市区町村長に提出され、そのほとんどの情報を含めた形で人口動態調査の死亡個票が作成され、保健所・都道府県を介して厚生労働省に集約される。厚生労働省ではこの死亡個票を元に WHO の定めたルールに則って一つの死因（原死因）を決定し、死亡統計が作成される。調査票の提出は従来紙媒体であったが、平成15年に電子化が始まり、現在ではほとんどの死亡個票が電子化されて蓄積されている。

死亡統計は行政やさまざまな分野の研究において重要な資料であるが、現代では一人が複数の疾患に罹患することも珍しくなく、死因を一つに決めることで捉えられなくなる情報があると考えられる。死亡診断書に記載される「死亡の原因」は死亡個票に含まれており、電子的に提出された分については研究利用可能となっている。個票について網羅的な分析は今井らによる 4 年分のデータに対する原死因決定の観点からの分析 [1]があるが、他にはほとんど行われていない。現在我々は利用可能な全データを用いて人口学的な観点から死因分析を行っており、その前処理として死亡原因の自由記載を病名コードに、期間を日数に自動で正規化した。本稿ではそこで得られた知見を報告する。

2. 目的

死亡診断書の原因・期間欄に対する自動正規化を実装し、正規化処理の観点からこの自由記載データの特徴を明らかにする。

3. 方法

3.1 解析対象

電子データとして利用可能な全データである 2003 から 2019 までの全死亡個票データを統計法に基づき申請し、解析対象とした。なお、個票データの二次利用となるため、本稿の結果は死亡統計の公表値とは異なる可能性がある。

3.2 自動正規化

正規化の対象となるのは「死亡の原因」のⅠ-ア、イ、ウ、エおよびⅡの 5 つの記入欄であり、それぞれについて原因と期間の2種類のフィールドがある。原因についてはICD-10コード、期間については単位を日数とする数値に正規化した。

3.2.1 原因の ICD-10 コード化

死亡原因の自動コーディングは、各欄の記載文字列を入力とし、そこ含まれる病名に ICD-10 コードを付与して出力する Named Entity Recognition (NER) とみなし、手法としては ICD-10 コードと対応づいた病名辞書による分割数最小法を用いた。概要を図 1 に示す。

辞書としてはICD-10対応標準病名マスター[2]（以下、マスター）から作成したもの、万病辞書[3]、独自の追加辞書を用いた。主として用いたものはマスターである。マスターは

ICD-10 コードだけでなく病態概念に該当する病名交換用コードも含み、定期的に更新が行われており、網羅性と信頼性のあるリソースと考えた。万病辞書はマスターに収載されていない病名表記を補完するため、追加辞書はこれらの辞書でもカバーされない語を補完するために用いた。なお、辞書には修飾語など病名以外の語も含めたが、ICD-10 コードの決定の際には無視している。

マスターは研究実施時点で公開されている全てのバージョン(ver.2.10～5.07)の全てのテーブルを統合して用いた。これは、解析対象データが最も古いもので平成 15 年のものであり、現在は廃止されている病名も出現する可能性があるためである。アップデートに伴いコードや修飾語の分類が変更されることがあるため、病名については病名管理番号と移行先病名管理番号、修飾語については修飾語交換用コードにより用語間を紐付け、各表記に対して最新のコード・分類を対応づけた。またその表記が収載されたバージョンの公開日・終了日(次のバージョンの公開日の前日)も対応付けた。削除フラグは考慮していない。

マスターの病名については、病名そのものの他に「・」を削除したもの、さらに「部」「型」「性」を削除したものを辞書に含めた。修飾語についても収載された表記の他に「の」「・」を削除したもの、さらに「、」を削除したものを辞書に含めた。

万病辞書はノイズとなることを避けるため、ICD-10 コードが付与されていない語を除いた。また、マスターを優先するために、マスターに収載されている語のみから構成され、かつその構成要素であるマスター病名と ICD-10 コードが同じであるものも除いた。

独自の辞書は、病名・修飾語辞書と、後述する前処理用のスペル訂正辞書の2種類を作成した。作成は、解析対象に自動正規化を適用し、病名を[D]にするといった抽象化を行った上で出現頻度順に並べ替え、比較的高頻度なものから目視で正規化できていないものをピックアップし、抽象化前の表現に立ち戻ることで辞書に追加すべき語をリストアップし、表記ゆれの範囲と考えられたものはスペル訂正辞書に、それ以外は病名辞書に加えた。病名辞書の収載語には可能な範囲で ICD-10 コード・病名交換用コード・マスターの修飾語分類(数字一桁)を付与した。

これらの辞書を作成する際には、前処理としてスペル訂正と、それに先立ってユニコード正規化や異体字の統一などの文字の正規化を行った。同じ前処理を解析対象に対しても適用した。

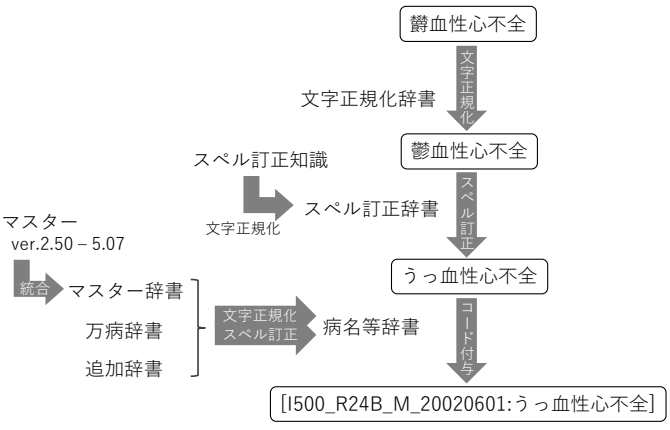

図 1. ICD-10 **コード化処理の概要**

以上の実装は全て形態素解析器 MeCab[4]を利用した。文字正規化・スペル訂正・病名辞書の 3 つのシステム辞書を作成し、パイプライン処理を行った。MeCab を用いた理由は実行速度であり、これにより試行錯誤による独自辞書の作成が行いやすくなった。

3.2.2 期間の日数化

期間の正規化については、病名のように辞書として利用可能なリソースが無く、また日付のような期間に相当する表現以外が見られたため、正規化のプログラムを作成した。

3.3 分析

3.2 で述べた自動正規化を年ごとに解析対象に適用し、以下を行った。

1. 原因欄フィールドの集計
   (ア) 原因欄の各フィールドを 1 単位として、1) 病名を 1 つ含む、2) 複数の病名を含む、3) 修飾語を 1 つ以上含み病名を含まない、4)「不詳」、5) その他の文字列を含む、6) 空欄、の 6 つに分類し集計した。
   (イ) 病名を 1 つ以上含むフィールドについて、「不詳」と「、」を削除し、出現要素の種類により1) 病名のみ、2) 病名・修飾語、3) 病名・修飾語・その他、4)病名・その他の 4 つに分類し集計した。
   (ウ) 複数病名を含むフィールド数を各欄について集計した。
   (エ) 空欄のフィールド数を各欄について集計した。
2. 病名出典の集計:出現した病名の出典を集計した。複数の辞書に掲載されていた場合、マスター、万病辞書、追加辞書の優先順位をつけ、重複の無いように計数した。また、出典ごとに病名の出現頻度を計数した。
3. 廃止・未収載病名の集計: マスター病名のうち、死亡日時点のバージョンよりも前に廃止された病名(廃止病名)、および後に追加された病名(未収載病名)の出現頻度を計数した。
4. 修飾語の集計: 出現した修飾語について、マスターで付与されている修飾語区分ごとに出現頻度を計数した。
5. 日数化処理のエラー分析: 日数化処理の結果、マイナス、または 150 年以上の数値が出力されたケースについて目視で確認し、原因の分類を行った。

4. 結果

4.1 解析対象

17 年分 15725292 件の死亡個票データを用いた。年ごとのデータ件数を図 2 に示す。死亡個票の電子化は徐々に行われており、2003 年の 9 万件弱から 2019 年には 140 万件近くになっていた。

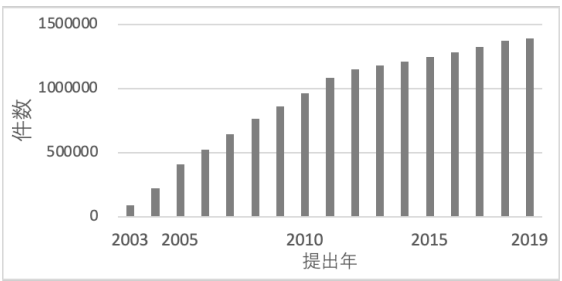

図 2 **解析対象の件数**

4.2 原因の ICD-10 コード化

マスターの統合の結果、55735 傷病名、2869 修飾語が得られた。万病辞書からは 90 の傷病名が追加された。追加辞書には 961 語が含まれていた。これらに前処理や記号削除などの処理を行い、74565 語から成る解析用辞書を得た。

万病辞書からの 90 語のうち ICD-10 の上一桁は R と T がやや多かった。

追加辞書に含まれる語は以下のようなものがあった。

- マスター病名の上位概念:「ノロウイルス感染症」
- マスター病名の表記ゆれ:「噴門部胃癌」(マスター病名「胃噴門癌」)
- コード化に寄与しない文字列:「おそらく」「推定」
- 内因死:「心臓死」
- 外因死:「縊死」
- 不明:「不詳」
- 物の名前:「餅」
- 薬剤名:「降圧薬」
- 場所、病院、地名:「風呂」

スペル訂正辞書に含まれる語は訂正の前後で表記ゆれの関係にあるが、以下のような種類があった。

- 字種の揺れ:「たこつぼ」-「蛸壺」
- 同義語:「MRSA」-「メチシリン耐性黄色ブドウ球菌」
- 読みが似ているもの:「出血」-「出欠」
- 形が似ているもの:「右」-「石」

スペル訂正辞書の中で特に表記ゆれの種類が多かったのは「アスペルギルス」等のカタカナ語、「僧帽弁」や「類天疱瘡」といった漢字表記であり、これらにおける表記ゆれの種類としては形が似ているものと読みが似ているものの混在であった。

4.3 期間の日数化

期間欄に記載される表現は「3 年」などの期間の他に発症時として日付や年齢があった。そのため、入力として期間欄の文字列だけでなく、発症時から期間を算出するために死亡日、死亡時の年齢または生年月日が必要であった。

原因欄に複数の傷病名が記載されている場合、期間欄には最大でその傷病名と同じ数の期間が記載される可能性がある。また原因欄に「①」「ア」など箇条書き記号が含まれ、それを期間欄で参照するケースも見られた。これらの情報を期間欄の処理時に手がかりとして用いるために原因欄の ICD-10 コード化の結果も入力として用いた。

アルゴリズムは大きく分けて 4 段階から成り、1) 文字の正規化、2) 文字列から構造化データの配列への変換、3) 構造化データから期間データへの変換、4) 期間データから日数を単位とする数値への変換であった。

処理の主要部分は 2 段階目の文字列から構造化データの配列への変換で、有限状態機械を実装した。原因欄のように分割数最小法としなかった理由は、上述のように原因欄の情報を手がかりとして利用するためである。

4.3 分析結果

表 1 に原因欄フィールドの集計を示す。およそ 3 分の 2 は空欄または「不詳」、3 分の 1 は病名を含むものであった。割合としては小さいが、病名を含まないが修飾語を含むフィールドが 0.5%程度存在した。単独の修飾語のみから成る記載例としては、「加齢」「脳梗塞後」「急性増悪」「十二指腸浸潤」、複数の修飾語のみから成る記載例としては「骨髄不全」「内臓破裂」「内因性疾患」「脳梗塞後後遺症」などがあった。空欄の割合は年とともに増える傾向にあった。

表 2 に病名を複数含むフィールドの集計、表 3 に空欄の集計を示す。死亡診断書記入マニュアル[5]に従えば 1 つの欄に複数の病名が記載されうるのは Ⅰ-エ欄および Ⅱ 欄であり、実際に最も多いのは全年通して Ⅱ 欄であったが、2 番目に多いのは I-ア欄であった。空欄については、必ず記入されるはずの Ⅰ-ア欄にも少数存在しており、例えば「誤嚥による窒息」などがあった。

表 1. 原因欄フィールドの集計

	病名	複数病名	修飾語	不詳	その他	空欄
2003	34.3%	0.7%	0.5%	2.9%	0.4%	61.2%
2004	34.2%	0.7%	0.5%	2.8%	0.4%	61.4%
2005	34.0%	0.7%	0.5%	2.8%	0.4%	61.6%
2006	33.9%	0.7%	0.5%	2.8%	0.4%	61.7%
2007	33.7%	0.7%	0.5%	2.8%	0.4%	61.9%
2008	33.5%	0.7%	0.5%	2.8%	0.4%	62.1%
2009	33.4%	0.7%	0.5%	2.7%	0.4%	62.4%
2010	33.1%	0.7%	0.5%	2.7%	0.4%	62.7%
2011	32.6%	0.7%	0.5%	2.6%	0.4%	63.3%
2012	32.7%	0.7%	0.5%	2.5%	0.4%	63.3%
2013	32.4%	0.7%	0.4%	2.5%	0.4%	63.6%
2014	32.2%	0.7%	0.5%	2.5%	0.4%	63.9%
2015	32.0%	0.6%	0.5%	2.4%	0.4%	64.2%
2016	31.9%	0.6%	0.5%	2.3%	0.4%	64.3%
2017	31.6%	0.6%	0.5%	2.3%	0.4%	64.7%
2018	31.4%	0.6%	0.4%	2.3%	0.4%	64.9%
2019	31.1%	0.6%	0.4%	2.1%	0.4%	65.4%

表 2. 複数病名フィールドの集計

	Ⅰ(ア)	Ⅰ(イ)	Ⅰ(ウ)	Ⅰ(エ)	Ⅱ
2003	14.2%	7.8%	1.6%	0.8%	75.6%
2004	14.1%	7.6%	1.6%	0.6%	76.1%
2005	14.2%	7.9%	1.5%	0.6%	75.9%
2006	13.6%	7.7%	1.7%	0.7%	76.4%
2007	13.1%	7.4%	1.7%	0.8%	77.1%
2008	12.8%	7.5%	1.6%	0.8%	77.4%
2009	12.8%	6.9%	1.6%	0.7%	78.0%
2010	12.3%	6.9%	1.6%	0.7%	78.6%
2011	12.6%	7.0%	1.4%	0.7%	78.3%
2012	12.4%	6.9%	1.4%	0.6%	78.6%
2013	11.8%	6.8%	1.5%	0.6%	79.2%
2014	12.2%	6.6%	1.5%	0.6%	79.1%
2015	11.9%	6.3%	1.3%	0.6%	79.9%
2016	11.6%	6.4%	1.3%	0.6%	80.1%
2017	11.7%	6.4%	1.3%	0.6%	80.0%
2018	11.5%	6.3%	1.3%	0.6%	80.3%
2019	11.9%	6.1%	1.2%	0.6%	80.2%

表 3. 空欄フィールドの集計

	Ⅰ(ア)	Ⅰ(イ)	Ⅰ(ウ)	Ⅰ(エ)	Ⅱ
2003	0.0%	17.8%	29.2%	32.0%	20.9%
2004	0.1%	18.0%	29.3%	31.9%	20.7%
2005	0.1%	18.1%	29.2%	31.9%	20.7%
2006	0.1%	18.2%	29.2%	31.8%	20.7%
2007	0.1%	18.4%	29.1%	31.7%	20.7%
2008	0.1%	18.6%	29.1%	31.6%	20.6%
2009	0.1%	18.8%	29.1%	31.5%	20.6%
2010	0.1%	19.0%	29.0%	31.3%	20.5%
2011	0.1%	19.3%	28.9%	31.1%	20.6%
2012	0.1%	19.3%	28.9%	31.1%	20.6%
2013	0.1%	19.5%	28.8%	30.9%	20.7%
2014	0.1%	19.6%	28.8%	30.8%	20.7%
2015	0.1%	19.8%	28.7%	30.7%	20.7%
2016	0.1%	19.9%	28.7%	30.6%	20.7%
2017	0.1%	20.0%	28.6%	30.5%	20.7%
2018	0.1%	20.1%	28.6%	30.4%	20.8%
2019	0.1%	20.4%	28.5%	30.2%	20.8%

表 4. 病名フィールドの集計

	病名のみ	＋修飾語	＋修飾語＋その他	＋その他
2003	87.9%	9.1%	1.6%	1.4%
2004	87.8%	9.2%	1.6%	1.4%
2005	88.1%	9.0%	1.5%	1.4%
2006	88.3%	8.9%	1.4%	1.3%
2007	88.4%	8.9%	1.4%	1.3%
2008	88.3%	9.0%	1.4%	1.3%
2009	88.3%	9.0%	1.4%	1.3%
2010	88.4%	9.0%	1.4%	1.2%
2011	88.5%	9.0%	1.3%	1.2%
2012	88.5%	9.1%	1.3%	1.1%
2013	88.5%	9.1%	1.3%	1.1%
2014	88.4%	9.2%	1.3%	1.1%
2015	88.3%	9.3%	1.3%	1.1%
2016	88.2%	9.4%	1.3%	1.1%
2017	88.0%	9.7%	1.3%	1.1%
2018	88.0%	9.7%	1.3%	1.0%
2019	88.0%	9.7%	1.2%	1.0%

表 5. 病名の出典の集計

	マスター	万病辞書	追加辞書
2003	96.5%	1.4%	2.1%
2004	96.7%	1.4%	2.0%
2005	96.8%	1.3%	1.9%
2006	96.7%	1.4%	1.9%
2007	96.6%	1.4%	2.0%
2008	96.7%	1.4%	1.9%
2009	96.6%	1.4%	2.0%
2010	96.7%	1.4%	1.9%
2011	96.6%	1.4%	2.0%
2012	96.7%	1.5%	1.8%
2013	96.7%	1.5%	1.8%
2014	96.8%	1.5%	1.7%
2015	96.8%	1.5%	1.7%
2016	96.9%	1.4%	1.6%
2017	97.0%	1.4%	1.6%
2018	97.0%	1.4%	1.5%
2019	97.1%	1.4%	1.5%

表 6. 削除病名、未収載病名の出現頻度

	削除病名	未収載病名
2003	115	21517
2004	2517	51782
2005	4201	92612
2006	6288	112220
2007	739	138233
2008	208	156236
2009	12	156884
2010	0	169905
2011	0	84273
2012	26	10824
2013	50	8741
2014	44	7866
2015	53	7575
2016	53	6545
2017	46	5245
2018	81	4187
2019	79	3846

表 7. 2019 年の出現頻度が高い語

	万病辞書	追加辞書:病名	追加辞書:その他
1	誤嚥	縊死	、
2	肺転移	蘇生後	)
3	虚血性心不全	低栄養	(
4	消化管穿孔	透析	の
5	心臓突然死	血液透析	推定
6	食物誤嚥	ペ-スメ-カ-	・
7	循環不全	循環器系疾患	による
8	嚥下機能障害	動脈閉塞	急性心臓死
9	急性虚血性心不全	転落	内因死
10	胸部大動脈解離	心機能不全	死

表 8. 修飾語の集計

	部位	位置	病因	経過表現	状態表現	患者帰属	その他	接尾語
2003	24.0%	23.5%	5.5%	5.9%	6.5%	2.4%	4.3%	27.8%
2004	11.0%	26.5%	6.8%	7.2%	7.5%	2.7%	5.2%	33.2%
2005	14.6%	25.2%	6.7%	7.2%	7.2%	1.6%	4.9%	32.5%
2006	19.0%	23.7%	6.5%	6.9%	6.9%	1.0%	4.8%	31.3%
2007	21.9%	23.0%	6.3%	6.7%	6.7%	0.8%	4.7%	30.0%
2008	21.1%	23.1%	6.3%	6.7%	6.7%	0.8%	4.6%	30.6%
2009	21.0%	23.3%	6.3%	6.6%	6.7%	0.7%	4.7%	30.7%
2010	20.9%	23.6%	6.2%	6.5%	6.7%	0.7%	4.6%	30.7%
2011	20.3%	23.8%	6.1%	6.7%	6.7%	0.7%	4.5%	31.1%
2012	20.1%	24.2%	6.1%	6.7%	6.6%	0.8%	4.5%	31.0%
2013	19.8%	24.7%	6.0%	6.7%	6.7%	0.8%	4.5%	30.7%
2014	19.6%	25.1%	5.9%	6.8%	6.7%	0.8%	4.5%	30.6%
2015	19.3%	25.7%	5.8%	6.7%	6.7%	0.8%	4.5%	30.5%
2016	19.1%	26.0%	5.7%	6.7%	6.7%	0.8%	4.5%	30.5%
2017	18.5%	26.1%	5.6%	7.2%	6.6%	0.8%	4.6%	30.7%
2018	18.2%	26.6%	5.4%	7.2%	6.5%	0.8%	4.6%	30.7%
2019	18.0%	27.2%	5.3%	7.1%	6.5%	0.8%	4.5%	30.5%

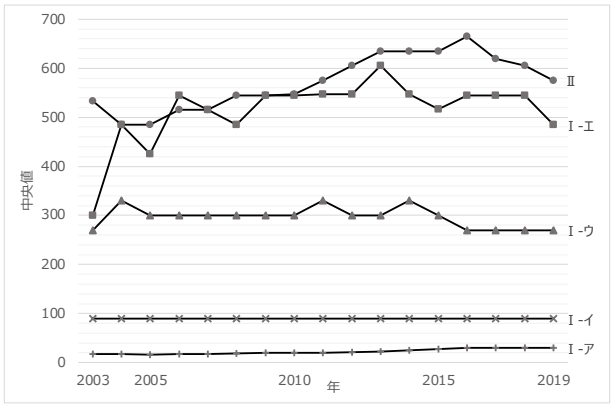

図 3. 日数の中央値

表 4 に病名を含むフィールドの構成要素による集計を示す。病名を含むフィールドのうち 88%程度が病名のみから構成されており、残りの 12%程度は修飾語やその他の文字列を含むものであった。この 12%は今回付与した ICD-10 コードが誤っている可能性がある。

表 5 に出現した病名の出典の集計を示す。病名はマスター病名が 97%を占めており、病名のリソースとしてマスターは十分適すると考えられた。

表 6 に削除病名・未収載病名の出現頻度を示す。削除病名の出現頻度は 2009 年以降は年間 100 件に満たず、この期

間のデータを対象とするならば対象データよりも過去のバージョンのマスターは使用せずとも解析結果にほとんど影響を及ぼさないと考えられた。削除病名に比べ未収載病名は数が多く、対象データよりも未来のバージョンのマスターも用いるほうがよいと考えられた。

表 7 に万病辞書と追加辞書の高頻度語上位 10 件を示す。万病辞書病名と追加辞書病名はそれぞれ 75 種、162 種が出現したが、うち 2019 年の出現回数が 50 件を超えていたものはそれぞれ 45 件、59 件であった。マスターの未収載病名は収載の前年に年間に数十～数百件出現しており、これらは今後マスターへ収載される可能性があると考えられた。追加辞書の病名以外の語のうち、因果関係を示す「による」が 4 位に挙がっているが、マニュアルによると 2 つの傷病名が因果関係にある場合は欄を分けるべきであり、死亡診断書の記載方法が十分に浸透していない可能性がある。他にも「経鼻栄養」など傷病名でなく原因欄の記載すべきでないものがあった。

表 8 に修飾語の区分ごとの出現割合を示す。修飾語のうち半分近くが部位と位置であり、次いで接尾語が多かった。

期間の日数化において、結果がマイナスとなるのはデータの誤りであり、期間欄の内容が死亡日より後の日付である場合が多かった。150 年以上となるのは多くは解析誤りだったが、解析困難な例として箇条書き記号の数字と連続して期間が記載されているものがあった。また「2 年から 10 年」といった幅が異様に広い表現が見られた。これは一つの傷病についての記載ではなく、対応する原因欄に複数の傷病名が記載されており、それらの最小値と最大値を記載したものと考えられる。

図 3 に日数の中央値を示す。特にⅠ-エとⅡ欄で振れ幅が大きく、解析の誤りの可能性が考えられた。解析ミスの原因としては桁あふれが考えられる。厚生労働省から提供されるデータでは各フィールドは固定長であり、原因についてはア-ウは 19 文字、エとⅡは 38 文字、期間についてはア-ウは 8 文字、エとⅡは 16 文字である。これを超えた記載がある場合は備考欄に続きを記入するものとされている。欄ごとに文字数を調べたところ、1%未満ではあるが最大文字数に達しているものが存在した。また原因・期間ともに内容的に不自然に途切れているものがあった。しかし備考欄は十分に構造化されているわけではなく、コード化・日数化の際に自動でこれを取り入れる方法は自明ではない。

5. 考察

コード化・日数化処理について、原因・期間ともにそのほとんどは今回のタスク設定で処理可能である。機械学習は入力が文でないため、NER として解くのであれば導入の意義は薄いと考えている。ただし、より正確な正規化処理に対してはその限りではなく、1) 備考欄を考慮にいれる、2) 1 つの原因欄に因果関係を持つ 2 つの傷病名が含まれていた場合はその関係も含めて認識する、3) 記載内容のうち「手段及び状況」など他の欄に記載すべき内容を認識する、といったことが必要である。つまり、最も正確にこの課題を扱うのならば、欄ごとの処理ではなく、全体を入力として、時間情報を付与した疾患と治療など外因の因果関係グラフを出力するようなタスクとして扱う必要がある。

死亡個票にはその作成過程から複数種類の『誤り』が混在する。作成過程とは、1) 医師・歯科医師が記載内容を決め、2) 医師・歯科医師がそれを記載し、3) 市区町村の担当者がそれを読み取り、4) 市区町村の担当者が調査票に転記または登録システムに入力する、さらに 5) 調査票を OCR で読み取る（平成 29 年度までで廃止）、というものである。段階 1 では本稿の結果にも現れたようにマニュアルに沿わない記載方法が取られうる。以降は表記ゆれが追加され、段階 2 では字種の揺れや同義語、段階 3 では形の似ている文字、段階 4 では読みの似ている語（変換ミス）、段階 5 では形の似ている文字へと表記が変わりうる。今回の ICD-10 コード化では段階 2 以降を対象としているが、スペル訂正辞書でこれらの分類情報は付与しておらず、今後の課題と考えている。段階 1 については、まず医師・歯科医師への教育が考えられるが、マニュアルは決して短いものではなく、作成時に自然と参照できるようなインストラクションが効果的と考える。電子的に作成するのであれば、入力欄の誤りのサジェストや病歴の自由入力・診療記録から各欄の入力内容の候補を提示するといった技術的な支援、さらに段階 2 の揺れに対応することも可能と考えられる。手書きで無くなることで上述の作成過程の段階 3 で発生する表記ゆれが、電子的に提出可能となれば段階 4 の変換ミスが無くなる。

期間については上記の段階 1 のみで揺れが発生していると考えられたが、表記そのものよりも傷病名との対応付けに困難があった。これについても電子的な入力環境のインターフェース次第で解析が容易になると考えられる。

マスターによる ICD-10 コード化の課題として、修飾語の扱いが挙げられる。Post-coordination によって ICD-10 コードが変わる可能性があるばかりでなく、修飾語は単独で用いても特定の疾患を示すものがあったため、修飾語の扱いを検討する必要があることがわかった。

6. 結論

2003 年から 2019 年までの死亡個票データの死亡の原因欄と期間欄を対象として傷病名の ICD-10 コード化・期間の日数化を適用した。死亡診断書に記載される傷病名のうち 97%程度はマスターでカバーされ、辞書によるシンプルな手法で 88%程度は確実なコード化が可能であることがわかった。またさまざまな表記揺れが存在しており、解消のためには死亡診断書の作成を、可能であれば提出も電子的に行うことが効果的と考えられた。

謝辞

本研究は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) JP20AB1001 の研究成果であり、本研究で使用した「人口動態調査」に関する分析結果には、統計法第 33 条の規定に基づき、調査票情報を二次利用したものが含まれている。

参考文献

1) 厚生労働科学研究費補助金 行政政策研究分野 政策科学総合研究（統計情報総合研究）「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」 令和元年度総括・分担研究報告書. 今井健, 2020.
2) ICD10 対応標準病名マスター [https://www2.medis.or.jp/stdcd/byomei/index.html (cited 2021-Aug-31)]
3) 万病辞書 [https://sociocom.naist.jp/manbyou-dic/ (cited 2021-Aug-31)]
4) MeCab: Yet Another Part-of-Speech and Morphological Analyzer [https://taku910.github.io/mecab/ (cited 2021-Aug-31)]
5) 死亡診断書（死体検案書）記入マニュアル 令和 3 年度版. 厚生労働省医政局政策統括官（統計・情報政策担当）, 2021. [https://www.mhlw.go.jp/toukei/manual/dl/manual_r03.pdf (cited 2021-Aug-27)].