原死因確定作業についての実態・問題点の把握，ならびに正確・効率性向上に向けた機械学習の適用可能性と課題に関する調査研究

今井（イマイ）　健（タケシ）＊1　明神（ミョウジン）　大也（トモヤ）＊2＊4
大井川（オオイガワ）　仁美（ヒトミ）＊3　香川（カガワ）　璃奈（リナ）＊6　今村（イマムラ）　知明（トモアキ）＊5

目的　人口動態調査は国勢調査と並ぶ国の主要統計で公衆衛生施策の中心的資料である。本研究ではこの中で死亡票に着目し，わが国における原死因確定作業の実態と問題点を明らかにすること，また，正確・効率性向上に向けた機械学習の適用可能性と課題について調査・検討を行うことを目的とした。

方法　まず，文献と厚生労働省担当者へのヒアリングを通じて，わが国における原死因確定プロセスの概要ならびに現状のオートコーディングシステムと人手確認作業の実態について調査を行った。次に，検証用に作成したダミーの死亡診断書ならびに原死因確定作業データを対象とし，諸外国で広く用いられているオートコーディングシステムIrisを用いた原死因確定作業の検証を行った。また，これらのプロセス分析結果を元に，正確・効率性向上のための機械学習の適用可能性について検討を行った。

結果　現行のオートコーディングシステムはルールに基づいた処理であること，同システムでICD-10コードが完全付与できないケースや付帯情報が含まれている場合については人手での目視確認による原死因確定作業が行われ，月約10万件の死亡票データのうち，約４割を占めていることが判明した。また死亡票ダミーデータを用いた分析から，特に付帯情報は最終原死因コードの確定に大きく影響を及ぼし得ることが判明した。原死因確定作業のさらなる高精度化と効率化のための機械学習の適用については，①Ⅰ欄Ⅱ欄傷病名に対するICD-10コーディングや②ICD-10コードの組み合わせからの選択ルールに基づいた仮原死因確定作業については，機械学習の適用が困難，あるいはメリットが大きくない一方で，③付帯情報を考慮した最終的な原死因確定プロセスについては，機械学習適用のメリットが大きいことが判明した。また，機械学習の適用シナリオとしては，「プロセス全体に対する適用」と「プロセス分解による部分的適用」の２つが考えられたが，後者は前者に比べて大幅な精度向上が期待できると共に，付帯情報の影響に特化して学習できる点からより適していると考えられた。

結論　今後，わが国の原死因確定プロセスに対し，上記のシナリオに準じ機械学習を部分適用することによって，現状の月約４万件に及ぶ人手確認作業の大幅な効率化と高精度化が期待できると考えられた。

キーワード　人口動態調査，死亡票，原死因確定プロセス，ICD-10，機械学習

＊1 東京大学大学院医学系研究科疾患生命工学センター准教授　＊2 奈良県立医科大学病理診断学講座医員
＊3 同MBT学講座博士課程　＊4 同公衆衛生学講座博士課程　＊5 同公衆衛生学講座教授
＊6 筑波大学医学医療系医療情報マネジメント学講師

Ⅰ 緒　　言

わが国において，人口動態調査は国勢調査と並ぶ国の主要統計であることから，統計法により基幹統計調査として位置づけられており，公衆衛生施策の中心的資料となっている。人口動態調査に関する届出は出生届・死亡届・死産届・婚姻届・離婚届の５種類ある。

人口動態調査死亡票（または死産票）は，死亡届（または死産届）に含まれる死亡診断書／死体検案書（または死産証書／死胎検案書）をもとに市区町村が作成し，さらにこの人口動態調査死亡票（または死産票）からICD-10コードに準拠したオートコーディングシステムと人手による確認作業を経て，原死因が確定されている。

オートコーディングシステムの仕組みは二段階構成となっており，①死亡票（または死産票）の「死亡の原因」Ⅰ欄とⅡ欄にある傷病名に対するICDコードの付与，②原死因の選択ルールに基づいた原死因の決定である。ただし，「死亡の原因」以外の付帯情報項目（死亡の種類，傷害の発生日時・場所，手段及び状況，あるいは「その他特に付言すべきことがら」等）に何らかの記載があったり，自動でのICDコード付与が困難であったり，原死因選択ルールに則らなかったりするという理由により，職員が手動で原死因確定作業を行わなければならないケースも多く存在している。

今後さらなる正確性向上と効率化のためには，近年発展目覚ましいAI関連技術，特に深層学習をはじめとする機械学習技術の援用による，原死因確定プロセスの支援手法の確立が重要であると考えられる。そこで本研究では，原死因確定作業の実態と問題点を明らかにすること，また，効率・正確性向上に向けた機械学習の適用可能性とそこにおける課題について調査・検討を行うことを目的とした。

なお，人口動態調査によると2017年度１年間の死亡人数は1,340,397人，死産件数は20,358件であったことから，本研究では件数の多い死亡届に着目した。

Ⅱ 方　　法

まず，文献調査と厚生労働省関係者へのヒアリングを元にして，わが国における原死因確定プロセスの調査を行った（**Ⅲ-（１）参照**）。調査に用いた文献は以下のとおりである。

①平成30年人口動態調査必携（厚生労働省政策統括官編）

②疾病，傷害及び死因の統計分類提要 ICD-10（2013年版）準拠 第二巻 総論

③ICDの改正に伴う人口動態死因オートコーディングシステム，人口動態データプロセッシングシステム等のシステム改修業務調達仕様書[1)]

①はわが国における人口動態調査の概要とどのような流れで調査が行われているかを把握するため，②はWHOが定める死因コーディングについてのルールおよびガイドラインの内容把握のため，③は厚生労働省内にて用いられているオートコーディングシステムならびに目視確認による処理過程の調査のために用いた。また，文献だけでは明らかにならなかった詳細については，厚生労働省関係者へのヒアリングにより調査を行った。

次に，検証用データを用いてオートコーディングシステムの挙動の調査を行った。厚生労働省内のオートコーディングシステムは利用することができないが，類似のものとして無償公開され欧米で広く用いられているオートコーディングツールIrisを利用した。Irisは日本語病名には対応していないが，Ⅰ欄Ⅱ欄傷病名のICD-10コード入力に対応しており，これを用いてオートコーディングの挙動をおおむね再現することができる。検証用データとしては，医師２名（病理診断医，整形外科医）の協力を得て臨床上の経験を元にダミーの死亡診断書18例を作成し，WHOが定めるルールに従って原死因コードを確定した。また，これとIrisによる出力結果との比較を行うことで，現状の原死因確定プロセスにおける自動処理部分の挙動につい

て分析を行った（Ⅲ−（２）参照）。

最後に，上記の結果を元に，原死因確定プロセスの支援における機械学習の適用可能性について検討を行った（Ⅳ参照）。

Ⅲ 結　　果

（１） 文献とヒアリング調査に基づく原死因確定プロセスの現状

人口動態調査に関する届け出は出生届・死亡届・死産届・婚姻届・離婚届の５種類あるが，これらの届を市町村が取りまとめて最終的に厚生労働省と法務省に提出される。このうち，死亡届は市区町村で死亡調査票になり，保健所・都道府県を経て厚生労働省に送付される。死亡届には死亡診断書（死体検案書）が含まれており，基本的な情報（氏名，生年月日，死亡日時，場所等），死亡の原因（Ⅰ欄Ⅱ欄傷病名），以外に各種の付帯情報（手術解剖の有無と所見，死因の種類，外因死の状況，生後１年未満死の追加事項，その他付言すべきことがら）などの情報が記載されている。

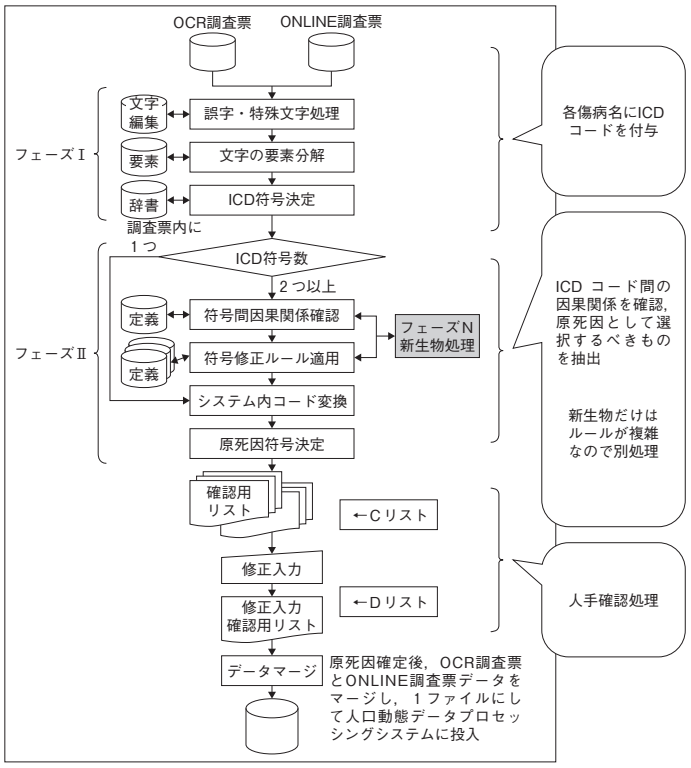
厚生労働省では送付されたこれらの死亡票データに対し，オートコーディングシステムにより，死亡票における死因（Ⅰ欄Ⅱ欄の各傷病名）のICD-10符号化，各種チェックリストの出力，データ修正および必要があれば人手による目視確認を経て，最終的な原死因コードを確定している。このプロセスの詳細は公開情報が極めて限定されているが，入手可能な文献資料とヒアリングの結果，以下のような処理手順であることが判明した（図１）。

1） フェーズⅠ処理（ICD-10コード付与）

フェーズⅠでは，誤字脱字処理と，死亡票のⅠ欄Ⅱ欄傷病名に対するICD-10コードの付与が行われる。死亡票データの収集についてはほぼ100％近く（98％）が電子的に収集されているが，中には死亡診断書からの転記時エラーやOCRの読み取りミスによる誤字脱字の存在（例：「肝不全」を「月干不全」とするなど）が存在する。このような事例に対し，一定の誤字脱字や特殊文字の処理を行っている。

死亡票のⅠ欄Ⅱ欄に記載された全傷病名に対するICD-10コードの付与処理については，システム内の辞書（病名の語幹と修飾語を含む）へのマッチングを行い，与えられた傷病名文字列を辞書内の形態素で被覆しようとするルールベースのアルゴリズムと推察された。この時，(A)完全被覆できICD-10コード付与可能なケース，(B)傷病名の一部のみマッチング可能でICD-10コードが仮付与されるケース，(C)全くの未知語でありICD-10コード付与不可能なケースが考えられるが，(A)(B)を合わせ「何らかのICD-10コードが付与／仮付与できるもの」が全体の約98

図１　原死因確定プロセスのフロー



注　「ICDの改正に伴う人口動態死因オートコーディングシステム，人口動態データプロセッシングシステム等のシステム改修業務調達仕様書」別添資料[1]より引用。一部改変

％ということであった。ただし，(B)の仮付与，あるいは(C)の付与不可能なケースはチェックフラグが立てられ，人手確認に回される。

2)　フェーズⅡ処理（原死因選択）

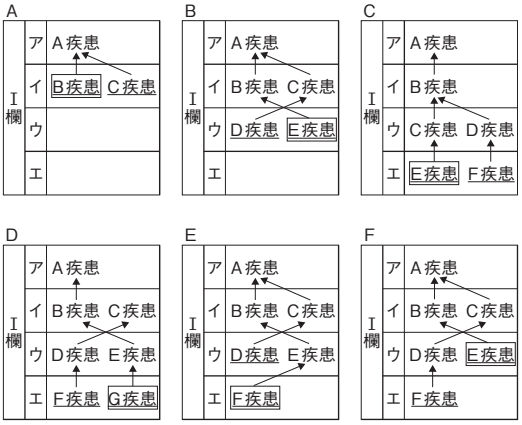
フェーズⅡでは，Ⅰ欄Ⅱ欄傷病名に対し付与されたICD-10コードの組み合わせから「疾病，傷害及び死因の統計分類提要 ICD-10（2013年版）準拠 第二巻 総論」（インストラクションマニュアル）に示されている原死因選択のためのルールならびに修正ルールが適用され，ルールベースで仮の原死因（最終確定ではない）が選択される。

WHOが定めた原死因選択の一般原則は，「死亡診断書に多数の病態が記載されている場合にはⅠ欄の最下欄に単独で記載された病態が，その上欄に記載されたすべての病態を引き起こす可能性がある場合に限り，その病態を選ぶ」とされている。ただし，この一般原則に従わない場合（因果関係が複数ある，因果関係が同定できないなど）には別途の選択ルールが存在し，上記「総論」には，このようなイレギュラーな事例について，多岐にわたり非常に細かく規則が示されている。一例として**図２**に「複雑な因果関係における原死因選択ルール」の例を示す。四角囲みのものが選択するべき原死因である。オートコーディングシステムでは，このような多くの規則を網羅的にルールテーブル化し，Ⅰ欄のアとイ，イとウ，ウとエの傷病名の因果関係のチェックなどを行い，原死因を確定しているということであった。しかし本処理で自動的に原死因が決定できない，あるいは疑義が残る場合にはフェーズⅠ同様，リジェクト／ワーニングコードが付与され，人手確認に回される。例えば「老衰」が年齢とあっていない，希少疾患であるなど多くのケースが存在する。

3)　付帯情報の有無チェック

さらに，「死亡の原因」Ⅰ欄Ⅱ欄傷病名の他に，何らかの付帯情報（手術解剖の有無と所見，外因死の追加事項，生後１年未満死の追加事項，その他付言すべきことがら等）の記述が存在する場合は，チェックフラグが立てられて，人手確認処理に回される。これは付帯情報の内容によっては，Ⅰ欄Ⅱ欄の傷病名以外に原死因が変更される可能性があるためである。何らかの付帯情報があるものは全体の約33％を占める。

図２　競合する因果関係がある場合の原死因選択例



注　「疾病，傷害及び死因の統計分類提要 ICD-10（2013年版）準拠 第二巻 総論」より引用

4)　原死因確定処理

月平均して約10.8万件の死亡票のうち，これまでの工程で，フェーズⅠにてすべての傷病名にICD-10コードが完全に付与され，フェーズⅡで仮の原死因選択が可能であり，また付帯情報が無い場合は，人手チェックが行われず自動的にシステムが選択したICD-10コードにて原死因が確定される（約6.8万件／月・64％）。それ以外は，傷病名にICD-10コードが付与できない，原死因が選択できないか疑義がある，何らかの付帯情報が存在するという原因によりCリスト，Dリストと呼ばれる確認用リストとして出力され，職員による目視確認を経たのちに最終的な原死因が確定される（約４万件／月・36％）。

また，付帯情報を人手で確認し原死因が変わるケースとしては，例えばⅠ欄(ア)に肺炎，手術欄に「胃悪性腫瘍切除術・１週間前」とある場合は，本来Ⅰ欄(イ)に胃癌と書くべきとみなしこれを原死因として選択する。あるいはⅠ欄(ア)で損傷の記述があっても，手段・状況欄で自殺，飛び降り，あるいは交通事故とわかるとそれを外因として優先する。また，病名に「細菌性肺炎」とあるが解剖欄の情報で菌の種類が

わかる場合は，詳細化してICD-10コードを付与するなど多様な事例が存在することが判明した。

ただし，この過程では「確認後得られた最終的な原死因コード」のみが電子的に記録されており，その他の中間情報（どの付帯情報とどの病名をあわせて考慮したか，など判断過程に関する情報，ICD-10コードが自動付与できない病名に対する正しいコード，入力データの誤りに対する修正結果など）は電子的に記録されていない。

（２） 死亡診断書例とIrisを用いた原死因確定プロセスの検討

わが国では，人口動態死因オートコーディングシステムを独自に開発・運用しているが，世界では近年，“Iris”とよばれるオートコーディングシステムを利用する国が増えており，現在ヨーロッパを中心に約20カ国で使用されている。開発は仏，独，伊，ハンガリー，米の５カ国の保健機関を中心としたIRIS INSTITUTEであり2008年から運用が開始されている。Irisは日本語をサポートしていないが，Ⅰ欄Ⅱ欄傷病名をICD-10コーディングした結果を入力すれば，原死因選択ルール処理の動作を確認することは可能である。今回，検証用データとして作成したダミーの死亡診断書18例を用い，付帯情報を除いた病名部分をICD-10コード化し，Irisに入力してオートコーディング処理の検証を行った。結果，18例中で12例は人手による原死因確定結果と一致，残りの６例のうち，入力から付帯情報を除外したため判断材料が足らずにリジェクト（“Not to be used for underlying cause（see volume 2 chapter 4.2.5)”）とされたものが４例，Irisにより原死因が確定されたものの，付帯情報を考慮すると人手による原死因確定結果と異なるものが２例という結果となった。

検証データが少ないという限界はあるが，付帯情報が影響を与えない場合については，WHOの原死因選択ルールセットに基づいて，Irisにて適切に原死因が確定されることが確認された。また，付帯情報が影響を与えた６例については，付帯情報の中身を加味した上で原死因コードの人手修正が必要であることも確認された。この付帯情報には「２月５日自宅で高所のものを取ろうとしていすから誤って転落し，受傷。四肢麻痺となり，救急要請した」といった状況説明など，自然言語文章の意味を解釈しないと正しくコーディングできないケースも多く，この処理が現状のオートコーディングシステムの自動処理では実現できていない大きな課題であることが判明した。

Ⅳ 考　　察

本節では，機械学習の適用可能性について考察する。Ⅲ-（１）より，原死因確定プロセスでは，まずフェーズⅠのⅠ欄Ⅱ欄傷病名に対するICD-10コード付与とフェーズⅡのルールベースによる原死因選択処理を経て問題なく自動で仮原死因が決定され，かつ付帯情報が無いもの（約６割）についてはオートコーディングシステムの出力結果が採用されている。一方，フェーズⅠ，Ⅱで問題がある，あるいは何らかの付帯情報があるもの（約４割）については，人手確認により最終原死因確定処理が行われている。従って機械学習を適用することによって原死因確定プロセスの高精度化・効率化を図る場合は，後者に対する支援が必要であるが，それには以下のような要素が考えられる。

（１） 各病名に対するICD-10コード付与

フェーズⅠで，Ⅰ欄Ⅱ欄傷病名に対しICD-10コードが自動的に付与できない事例への支援であるが，これは自動ICD-10コーディングとして良く知られているタスクであり古くから多くの研究がなされている。これまでの手法は大きく分けて㈀ICD-10のカテゴリ分類に関する知識（ルール）を計算機処理可能な形式で記述し，これに基づいて自動分類を行うもの（知識ベース[2)-4)]）と㈁ICD-10コードが既知である疾患名との類似度計算など統計的手法によりICD-10コードを自動付与するもの（用例ベース[5)6)]）が存在し，わが国でも知識ベース[7)]，用例ベー

ス[8]双方の研究が行われている。知識ベース手法の利点は，件数が少ない事例についてもルール化により自動コード付与が可能であること，欠点は知識ベースを構築するコストがかかることである。用例ベース手法の利点は知識ベース構築に関するコストが削減できること，欠点は既知の用例が存在しないと自動コード付与ができないこと，また，頻度に依存するため用例が少ないコードへは正しくコーディングしづらいことである。

現在のオートコーディングシステムは知識ベースの手法を用いている。これを昨今進展著しい機械学習技術を用いて用例ベースの手法で再構築した場合，年間130万件の死亡票データが存在し頻度情報は十分に得られるため，欠点である「用例が少ないコードへの分類問題」はある程度解消されることが予想される。しかしながらヒアリングによると，ICD-10コードが自動付与できない傷病名があるために人手確認を行った場合も，正しいICD-10コードは記録されておらず，電子的に保存されているのはあくまで「最終的な原死因確定結果」のみということであった。従って死亡票データから直接的に学習に必要な教師データを得ることができず，別途未コード化傷病名に対する膨大なコード付与作業が必要となるため，ICD-10コード付与プロセスについて機械学習を適用することは難しいと考えられた。

（２） 原死因の選択処理

フェーズⅡでは，WHOが定めたルールに基づき原死因選択処理が行われる。Ⅲ-（２）の結果から，厚生労働省のオートコーディングシステムと同様に，一般公開されているIrisでもこれらのルール処理が実現されていることが確認できた。仮に，人手確認に回らず原死因確定に至った約６割のデータを用いれば，「ICD-10コードの組み合わせから最終的な原死因コードを選択するルールの総体」を機械学習モデルで置換することも可能と考えられるが，既に実現されている処理を再学習するメリットは薄い。むしろ現状のIrisは複数国が参加して開発・メンテナンスが行われていることから，今後のルールテーブルのメンテナンスコストやICD-11導入の際の更新コストを考えると，Irisの積極的な活用は有力な選択肢と考えられる。

一方，この過程で人手確認が必要なのは選択ルールでも原死因が決定できない，あるいは疑義がある場合である。しかし，付帯情報の影響ではなく，純粋に原死因選択ルールに起因してリジェクト・ワーニング出力されるものは非常に数が少なく，対処がケースバイケースである。例えば，極めて稀な疾患である，あるいは生年月日や死亡日時の入力ミスにより「老衰」が年齢と合わないという場合は人手確認が必要であり，自動的に対処を行うことは困難である。従ってこれらを支援する際に機械学習は適さない。むしろこのようなケースこそ人手確認するべき事例であると考えられる。

（３） 付帯情報を考慮した最終的な原死因確定プロセス

何らかの付帯情報が含まれる場合は，必ず人手確認が行われる。これは月に約４万件行われる人手確認の理由の大多数を占め，機械学習適用による支援は極めてメリットが大きい。ここでは，手術，解剖欄，手段及び状況，付言欄の自然言語文章の解釈を加味した上での最終的な原死因確定が大きな課題である。仮に統計法33条に基づき死亡票データの提供を受けたとしても，オートコーディングシステムの内部処理過程データ（中間生成物）を入手することはできない。従ってこのプロセスに対し機械学習を行う際には，入手可能な(A)Ⅰ欄Ⅱ欄の傷病名の組み合わせ，(B)各種付帯情報，(C)最終的に確定された（修正された場合を含む）原死因コードのみで行う必要がある。ここから以下の２つの機械学習適用ストーリーが考えられた。

ストーリー１：プロセス全体に対する機械学習の適用

上記の(A)(B)の全情報を入力として，(C)を出力とするペアを学習用教師データとして機械学習を適用するパターンである。現行のオートコー

ディングシステムは(A)についてICD-10コードを完全付与／部分付与／付与できないのケースがあり，この中間生成物は取得することが困難である。従ってこの学習にICD-10コードを用いることはできない。また，ICD-10コードの組み合わせによる原死因選択ルールも用いることができず，付帯情報がⅠ欄Ⅱ欄のどの傷病名について述べたものであるか，という情報も不明である。しかしながら十分な件数があることから(A)(B)から直接(C)を導き出す深層学習等の適用により，ある程度の精度で学習が行える可能性がある。実際には(A)(B)に対する分散表現をword2vecなど別な機械学習により事前に獲得しておき，これらを入力ベクトルとし，(C)の約14,000次元のベクトルを出力ベクトルとするニューラルネットワークを訓練するなどの手法が考えられる。全体の約６割は既存のオートコーディングシステムでも自動処理が可能な対象（確定的に原死因が決定できるケース）であり，これらを含むことによって，真に学習するべき対象（人手による確認が必要なケース）の相対件数が減ってしまうことによって，頻度が多いケースに引きずられ，うまく学習できない可能性が考えられる。また(B)の付帯情報は患者の非常に多様な背景情報（産後３カ月であった，あるいは傷害が生じた状況など）が含まれ，その他のパラメータも非常に多いことも考えると，次に述べるストーリー２よりも大幅に精度が落ちる可能性があることには留意するべきである。

ストーリー２：プロセスの分解による部分的機械学習の適用

プロセス全体のうち，現行のオートコーディングシステム，あるいはIrisの利用により確定的に決定できる部分は機械学習の対象とせず，人手による解釈が必要とされる部分にのみ注力して機械学習を適用するパターンである。日本語環境であっても，仮にⅠ欄Ⅱ欄の全傷病名に対してICD-10コーディングが可能であれば，Irisを用いて仮原死因（付帯情報を考慮しない原死因選択ルール適用結果）を定めることが可能である。従って，このようなケースに対象を絞ることによって，あとは仮の原死因が付帯情報によってどのように最終的な原死因コードへ変化するのか，あるいはしないのかという点のみに特化して学習することが可能である。

問題は全病名に対してICD-10コーディングが行えていることを仮定していることである。前述のとおり，ICD-10コーディングの学習を死亡票データのみによって行うことは不可能であることから，これに代わる方法としてICD-10対応電子カルテ用標準病名マスター（以下，標準病名マスター）の利用が考えられる。標準病名マスターは厚生労働省標準規格の１つであり，23,000を超える傷病名数，94,000を超える索引語を含み，約2,000の修飾語との組み合わせで，病名を表現することが可能で，またICD-10コードに紐付けられている。電子カルテへの搭載が進んでいるが，死亡診断書などの自由記載病名においては必ずしも用いられていない。死亡票データに対し，この標準病名マスターを用いてICD-10コードが完全付与できたものを選択して対象とすることで，Ⅰ欄Ⅱ欄傷病名にICD-10コードが完全付与された状況で付帯情報の影響のみに特化して学習することが可能となる。この場合に使用できる情報は以下のとおりである。

(A) Ⅰ欄Ⅱ欄の傷病名の組み合わせ
　(A-2)(A)に対する完全付与ICD-10コード
　(A-3) Irisに(A-2)を入力して得られる仮原死因コード
(B)各種付帯情報
(C)最終的に確定された（修正された場合を含む）原死因コード

ストーリー１に比較して，(A-2)(A-3)を用いることができること，また，(B)付帯情報が(A-3)の仮原死因コードに影響を与えるかどうかに学習を特化させることができるため，より精緻な学習ができると考えられる。「影響を与える場合，最終原死因コードは何に変わるか」まで学習できれば理想であるが，仮に「影響を与えるか否か」の２値分類を高精度で学習できるだけでも人手確認作業を大幅な削減することができるため極めて有用である。標準病名マスターでICD-10コードが完全付与できたケース

に絞ることでストーリー１よりも学習に使用できる件数が減ることが欠点であるが，死亡票データは十分な量があるため，さほど影響が大きくないと考えられる。

Ⅴ 結　　語

本研究では，死亡票に着目し，原死因確定作業の実態と問題点を明らかにすること，また，正確性向上に向けた機械学習の適用可能性と課題について調査・検討を行うことを目的とした。文献調査とヒアリングの結果，現行のオートコーディングシステムはルールベースの処理であること，同システムでICD-10コードが完全付与できないケースや付帯情報が含まれている場合については，人手での目視確認による原死因確定作業が行われ，月約10万件の死亡票データのうち約４割を占めていることが判明した。また，検証用の死亡票ダミーデータを用いた分析から，付帯情報，特に自然言語文の解釈が最終原死因コードの確定に大きく影響を及ぼし得ることが判明した。

原死因確定作業のさらなる高精度化と効率化のための機械学習の適用については，調査の結果，①Ⅰ欄Ⅱ欄傷病名に対するICD-10コーディングや②ICD-10コードの組み合わせからの選択ルールに基づいた仮原死因確定作業については機械学習の適用が困難，あるいはメリットが大きくない一方で，③付帯情報を考慮した最終的な原死因確定プロセスについては，機械学習適用のメリットが大きいことが判明した。また，③について機械学習を適用するストーリーとしては，「(A)プロセス全体に対する機械学習の適用」と「(B)プロセス分解による部分的機械学習の適用」の２つが考えられた。後者は，標準病名マスターですべての傷病名についてICD-10コードが完全付与された事例に限定し，Irisで仮原死因コードを決定した後に，付帯情報による原死因コードの変更の有無を学習させるものである。前者に比べて大幅な精度向上が期待できると共に，主に人手での確認作業が必要とされる付帯情報の影響に特化して学習できる点からより良いと考えられた。今後，このストーリーに準じて昨今進展著しい深層学習を始めとした機械学習を適用することによって，原死因確定における月約４万件に及ぶ人手確認作業の大幅な効率化と高精度化が期待できると考えられる。

本研究の内容は，厚生労働統計協会の平成30年度調査研究委託事業（主任研究者：今井健）に基づいている。

文　　献

1）ICDの改正に伴う人口動態死因オートコーディングシステム，人口動態データプロセッシングシステム等のシステム改修業務調達仕様書.（https://www.mhlw.go.jp/sinsei/chotatu/chotatu/kankeibunsho/20150831-1/index.html）2019.11.1.

2）Fabry P, Baud R, Ruch P, et al. A framebased representation of ICD-10. Stud Health Technol Inform. 2003；95：433-8.

3）Héja G, Surja'n G, Luka'csy G, et al. GALEN based formal representation of ICD10. Int J Med Inform. 2007 Feb-Mar；76(２-３)：118-23.

4）Jiang G, Pathak J, Chute CG. Formalizing ICD coding rules using Formal Concept Analysis. J Biomed Inform. 2009 Jun；42(３)：504-17.

5）Pakhomov SV, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. J Am Med Inform Assoc. 2006 Sep-Oct；13(５)：516-25.

6）Tagliabue G, Maghini A, Fabiano S, et al. Consistency and accuracy of diagnostic cancer codes generated by automated registration：comparison with manual registration. Popul Health Metr 2006；４：10.

7）Imai T, Kajino M, Sato M, et al. Development of structured ICD-10 and its application to computer-assisted ICD coding. Stud Health Technol Inform. 2010；160(Pt２)：1080-4.

8）Aramaki E, Imai T, Kajino M, et al. Statistical selector of the best multiple ICD-coding method. Stud Health Technol Inform. 2007；129(Pt１)：645-9.