別添資料２　Iris による仮原死因付与処理

はじめに

本年度新たに作成した死亡票と死亡個票の突合データから、ランダムサンプリングで抽出した50万件のデータを対象に、今後の300～500万件規模の研究のための実験を行った。具体的には、現行の人口動態死因オートコーディングシステム（以下、オートコーディングシステム）に替わり Iris を用いて原死因を付与し、その結果（以下、仮原死因）とオートコーディングシステムが導出した確定原死因との比較を行った。また、付帯情報の有無に関する情報も抽出し、付帯情報と追加コードの影響について考察を行った。

実験

方法

手順1.　対象データを抽出

本年度新たに作成した死亡票と死亡個票の突合データから、ランダムサンプリングで抽出した 50 万件に対し、各突合データ中の傷病名に全て ICD-10 コードが付与できたもの（320,112 件）を対象とする。

手順2.　対象データの確定原死因、追加コード、付帯情報有無の情報抽出

手順 1.で抽出したデータに関して、確定原死因と追加コードを抽出する。追加コードとは、死亡票における外因符号と母側符号のことを指す。また、付帯情報の有無を 0（付帯情報なし）と 1（付帯情報あり）で表現し抽出する。「付帯情報あり」の定義は、別添資料１「原死因確定プロセス調査（昨年度のアップデート）」の末尾 **参考資料「付帯情報まとめ」**の 4 列目に示す通りである。

手順3.　対象データを Iris 用のフォーマットに整形

突合データから Iris に必要な以下の情報を抽出し、Iris のフォーマットに則り整形を行う。

- 死亡票識別番号
- 生年月日
- 死亡年月日
- 性別
- I・II 欄の傷病名

Iris は Microsoft Access 形式であり、基本情報が含まれるテーブル“Ident”と I・II 欄に記載される傷病名の情報が記載されるテーブル“MedCod”の

2 つから構成される。そのため、Access で読み取り可能な tsv 形式ではじめにファイルの整形を行い、自動的に Access ファイルを生成するバッチファイルを使用することで Iris に適用可能な Access ファイルを生成する。ファイルの生成具体的な Iris のフォーマットは、**本資料末尾「<参考資料> Iris のフォーマット」**に示す。

また、これまでの研究により、表 1 左から 1 列目に示す ICD-10 コードは日本（標準病名マスター）独自のコードであるため、Iris がコーディングエラーを引き起こす。そのため、左から 2 列目のように、事前に ICD-10 コードの変更を行う。

表 1　日本独自コードと Iris の対応

対象 ICD-10 コード	変更後 ICD-10 コード	傷病名
J9699	J969	呼吸不全
J9609	J960	急性呼吸不全
E14	E149	糖尿病
J9619	J961	慢性呼吸不全
G903	G239	多系統萎縮症
E11	E119	2 型糖尿病
B59	J173	ニューモシスチス肺炎
I7020	I702	末梢動脈硬化症
J9691	J969	2 型呼吸不全
E10	E109	1 型糖尿病
M7265	M726	外陰部壊死性筋膜炎
R17	R198	黄疸
J9611	J961	慢性 2 型呼吸不全
E13	E139	B 型インスリン受容体異常症
J9601	J961	急性 2 型呼吸不全

手順4.　Iris による仮原死因付与

実験で用いる Iris のバージョンは Iris Version 5.6.0-Y2019S1 であり、原死因コーディング部分を行う MUSE のバージョンは MUSE 2.7 である。仮原死因の付与では、Iris のバッチ処理機能を用いて一括で原死因付与を行う。

手順5.　仮原死因と確定原死因の比較

Iris による仮原死因と、オートコーディングシステムによる確定原死因

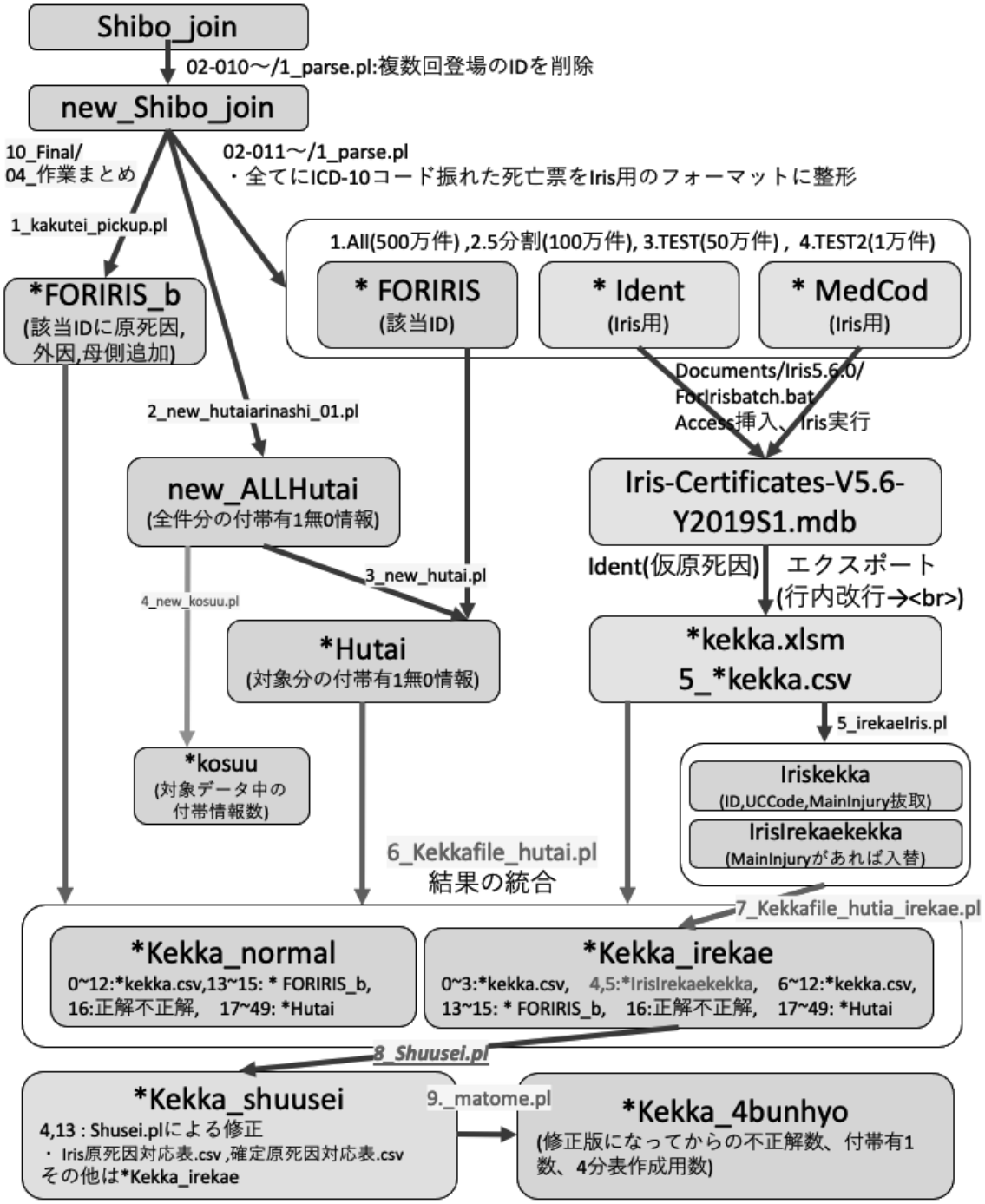
との比較を行い、一致率を算出する。ただし、比較を行う前に以下ことを行う。

- Iris の外因（複合）コードを仮原死因に変更
  - 厚生労働省へのヒアリングの結果、日本のオートコーディングシステムでは、S と T から始まるコードが原死因になり得るとし、死因統計を行う際、目視確認で付与する外因符号と合わせて集計を行っている。一方、WHO には、S と T に関するコードは原死因に使用しないという指針があり、Iris もそのルールに則っていることがわかった。そのため、Iris の仮原死因欄には、日本における外因コード、外因（複合）コード欄には日本における仮原死因が選ばれており、そのままの比較では一致しない。そこで、Iris で外因（複合）コードに記載がある場合は、仮原死因と交換する作業を行う。
- 標準病名マスターと Iris どちらにも存在しないが確定原死因に用いられる日本独自コードの対策
  - 厚生労働省へのヒアリングの結果、日本独自の詳細分類や、準拠する ICD-10 コードの違い(平成 28 年度以前は ICD-10 2003 年版、平成 29 年以降は 2013 年版）から、仮原死因と確定原死因の不一致が生じる場合があることがわかった。そこで、それぞれに関して、修正するか比較対象としない選択を行う。その基準は以下の通りである。
    - ✧ オートコーディングシステムによる確定原死因側
      - 1 桁多いため、仮原死因と一致しない
        - 桁を落とす
      - Iris の辞書に存在しない
        - 比較対象としない
      - 準拠する ICD-10 コードが変化した
        - 比較対象としない
    - ✧ Iris による仮原死因側
      - 1 桁多いため、確定原死因と一致しない
        - 桁を落とす

また、仮原死因と確定原死因が異なる場合は、仮原死因の変更があったとみなし、付帯情報と追加コードの有無を考慮して表を作成する。

以上の手順を自動化可能にするプログラムを作成した（次ページ Figure1 参照)。

(Figure. 1) 生成ファイルと処理プログラム概要



結果

50 万件のランダムサンプリングした突合データから、全ての病名に対して ICD-10 コードが付与できた 320,112 件に対し、Iris の処理時間は約 19 時間であった。確定原死因と仮原死因が比較可能であったのは、320,008 件であった。

仮原死因と確定原死因の一致率は、91.0%であった。また、仮原死因の変更有無と追加コードの変更有無、付帯情報の有無の表を表 2 に示す。

表 2 仮原死因変更・付帯情報・追加コードの有無別の件数

	付帯情報あり	**付帯情報なし**
仮原死因変更あり	**11,857(/320,008≒3.7%)**	**16,957(/320,008≒5.3%)**
追加コードあり	1,554(/81,695≒1.9%)	389(/238,313≒0.2%)
追加コードなし	10,303(/81,695≒12.6%)	16,568(/238,313≒7.0%)
仮原死因変更なし	**69,838(/320,008≒21.8%)**	**221,356(/320,008≒69.2%)**
追加コードあり	3,547(/81,695≒4.3%)	715(/238,313≒0.3%)
追加コードなし	66,291(/81,695≒81.1%)	220,641(/238,313≒92.6%)

考察

処理時間に関して、これまでの実験より、3,267 件に対する処理時間は約 4 分、6,484 件のデータに対する Iris の処理時間は約 12 分であった。6,484 件の場合の処理時間を基準に処理時間は件数の増加と等倍になると推測していたが、今回の 320,112 件に対する処理時間は推測の約 2 倍の時間を要していることがわかった。このことから、Iris では死亡票の件数が多いほど処理時間が爆発的に増えると考えられる。今後の研究では、対象データを 300 万～500 万件に拡張するが、全件を一括処理させるのではなく、多数に分割して一括処理を行うべきであり、分割の数を検討する必要がある。

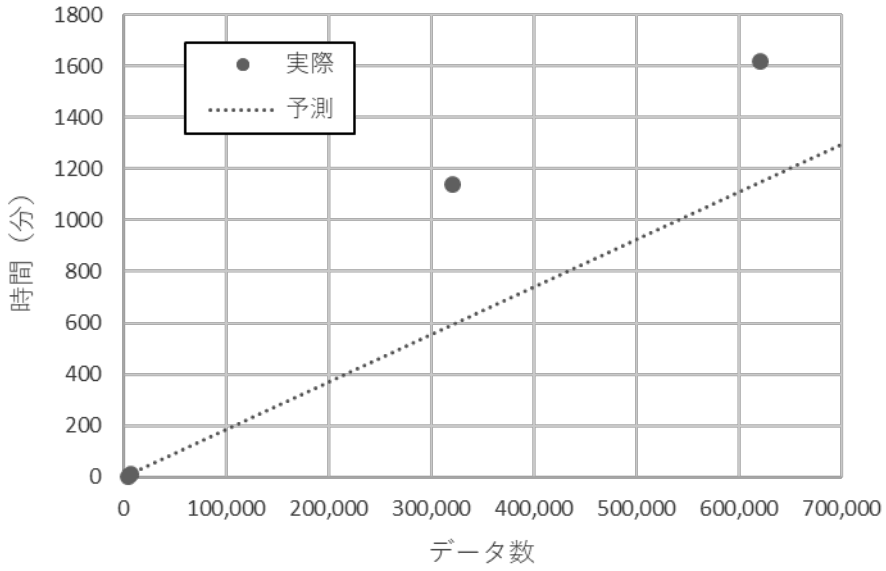


Figure 2　Iris の処理時間

確定原死因と仮原死因が比較に関して、手順 5.の操作により、機械学習に用いることのできるデータ数の減少が懸念されたが、比較対象とならないものはほぼ存在しないことがわかった。そのため、手順 5.の操作は、今後の研究でも用いていく。

仮原死因と確定原死因の一致率より、仮原死因の変更があるとみなされるものは全体の約 10%であることがわかった。ここで、研究班の研究対象は「付帯情報あり」のものであるため、表 2 の黄色部分（81,695 件）に限定すると、原死因の変更が行われるのは約 15%であることがわかった。また、原死因の変更以外にも、目視確認時に追加コードが付与されていることは、本年度行った原死因プロセスに関する研究で明らかになっている（別添資料 1 参照)。表 2 の追加コードに関する情報から、仮原死因の変更と追加コードの付与どちらも必要もないのは約 8 割であり、残りの 2 割は何らかの処理(原死因コードの変更・追加コードの付与)が必要であることになる。以上のことから、今後の研究として、付帯情報のある死亡票を対象に機械学習を用いて解くべき課題は、以下の 2 段階であると考える。

①対処（原死因コードの変更、外因や母側病態コードの追加）の必要がない 8 割の除去

②細かな対処内容の分類

これらの効果に関して、①が達成されることにより、8 割の人手処理の削減が可能になると考えられる。できることが期待でき、2 段階目の処理によって残りの 2 割に対する人手作業の効率化が図ることができると期待できる。また、2 段階の先としては、具体的な原死因変更先の ICD-10 コードや追加コードの提示が考えられる。

まとめ

本年度新たに作成した死亡票と死亡個票の突合データから、ランダムサンプリングで抽出した 50 万件のデータを対象に、Iris による仮原死因とオートコーディングシステムによる確定原死因の比較を行った。その結果、何らかの対処が必要なもの（原死因コードの変更、外因や母側病態コードの追加）は約 2 割であることがわかった。そして、今後の研究課題として、以下の 2 段階を挙げた。

① 対処（原死因コードの変更、外因や母側病態コードの追加）の必要がない 8 割の除去
② 細かな対処内容の分類

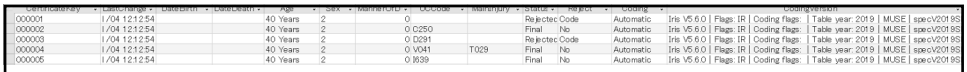
今後 300～500 万件の規模の研究のために留意すべき点として、Iris による仮原死因の付与の作業において、一括処理を行う死亡票の量を調整する必要がある。

<参考資料> Iris フォーマット

Irisに与えるファイル

- （理想）Access形式：2つのテーブルが必要
  - 末尾に”Ident”がつくテーブル
    - Ex. TestIdent
    - 死亡票の基本情報
  - 末尾に“MedCod”がつくテーブル
    - Ex. TestMedCod
    - 死亡票のⅠⅡ欄情報
- （方法）2種類のtsvをIrisのAccessファイルにインポート
  - ○○Ident.tsv
  - ○○MedCod.tsv

インポートだけで済むフォーマットをお願いしたい



Identテーブル用tsvフォーマット

列番号	項目	説明	対応
0	CertificateKey	識別番号（主キー）	数え上げ？ 市区町村コード等？
1	LastChange	最終更新	空で良い（後にIris更新）
2	DateBirth	生年月日	new_birthそのまま
3	DateDeath	死亡年月日	new_deathそのまま
4	Age	年齢	空で良い
5	Sex	性別	sex
6	MannerOfDeath	死亡の種類	0を記載
7	UCCode	原死因コード	空で良い（後にIris更新）
8	MainInjury	主傷病名	空で良い（後にIris更新）
9	Status	Irisの処理状況	Initialを記載（後にIris更新）
10	Reject	エラー理由	Noを記載（後にIris更新）
11	Coding	コーディング方法	Automaticを記載
12	CodingVersion	Irisバージョン	空で良い（後にIris更新）
13	CodingFlags	複数の候補がある	空で良い（後にIris更新）
14	SelectedCodes	ⅠⅡ欄読込結果	空で良い（後にIris更新）
15	SubstitutedCodes	ⅠⅡ欄Iris用の正規表現	空で良い（後にIris更新）
16	ErnCodes	現バージョンは関係なし	空で良い
17	AcmeCodes	MUSEへの読み込み	空で良い（後にIris更新）
18	MultipleCodes	MUSEの分析コード	空で良い（後にIris更新）
19	Comments	付言すべきことがら	空で良い
20	FreeText	備考	空で良い
21	ToDoList	Irisのリジェクト内容表示	空で良い（後にIris更新）
22	CoderReject	手動でリジェクト	空で良い

列番号	項目	説明	対応
23	DiagnosisModified	手動で修正	空で良い
24	Residence	国籍	空で良い
25	Name	氏名	空で良い
26	Address	住所	空で良い
27	AutopsyRequested	解剖有無	空で良い
28	AutopsyUsed	解剖主要所見	空で良い
29	RecentSurgery	手術有無	空で良い
30	DateOfSurgery	手術年月日	空で良い
31	ReasonSurgery	手術主要所見	空で良い
32	DateOfInjury	傷害発生年月日	空で良い
33	PlaceOfOccurrence	傷害発生場所	空で良い
34	ActivityCode	傷害発生場所コード	空で良い
35	ExternalFreeText	手段及び状況	空で良い
36	Pregnancy	妊娠状況	空で良い
37	PregnancyContributeDeath	死亡と妊娠の関与	空で良い
38	Stillbirth	死産有無	空で良い
39	MultiplePregnancy	単胎・多胎	空で良い
40	CompletedWeeks	妊娠週数	空で良い
41	BirthWeight	出生体重	空で良い
42	AgeOfMother	母の生年月日	空で良い
43	ConditionsMother	母体の状況	空で良い
44	CertImage	画像ファイル有無	空で良い

TestIdent.tsv変換規則

準備：0行目0列目～44列目にCertificateKey ～CertImage（前スライド参照）を記載

① 0列目(CertificateKey): 識別番号
- 一番上を0000001として数え上げ（桁数を揃える）

② 2列目(DateBirth)： 46(birth_ymd)

③ 3列目(DateDeath)： 49(death_ ymd)

④ 5列目(Sex)： 44(sex)

⑤ 6列目(MannerOfDeath)： 0

⑥ 9列目(Status)： Initial

⑦ 10列目(Reject)： No

⑧ 11列目(Coding)： Automatic

⑨ 他の欄は空

1死亡票につき1行

行列番号は0からカウントしています

MedCodテーブル用tsvフォーマット

列番号	項目	説明	対応
0	CertificateKey	識別番号（主キー）	数え上げ？ 市区町村コード等？
1	LineNb	傷病名記載箇所（主キー）	0-3がⅠ欄，5がⅡ欄
2	TextLine	傷病名	空で良い
3	CodeLine	ICD-10コード	ICD-10コード(期間)
4	IntervalLine	期間	空で良い（3列目に記載）
5	CodeOnly	ICD-10コードのみでのコーディング	1を記載
6	LineCoded	テキスト入力がある	1を記載（バッチ処理で結局1に変更される）

CertificateK	LineNb	TextLine	CodeLine	IntervalLine	CodeOnly	LineCoded
000001	0		G903		1	1
000001	1		J181		1	1
000002	0		K829		1	1
000002	1		K831		1	1
000002	2		C250		1	1
000003	0		I619		1	1
000003	1		I10		1	1
000003	2		N119		1	1
000003	3		D291		1	1

TestMedCod.tsv変換規則

準備： 0行目0列目～6列目にCertificateKey ～LineCodedを記載

※　List.txt の9(Ia_c) , 11(Ib_c) , 13(Ic_c) , 15(Id_c) ,17(II_c)に記載があれば，その都度，以下を生成．なければ次の死亡票へ

① 0列目(CertificateKey):　識別番号・・・ TestIdent.tsvと同じ

② 1列目(LineNb)：　ⅠⅡ欄の該当する番号

③ 2列目(TextLine)：　空

④ 3列目(CodeLine)：　③に対応するICD-10コード
- 香川先生により振られたICD-10コード
- 複数病名も記載可能（区切り文字任意）
  - Ex. C169,C189　とか C169 C189
- 期間がある場合は(数字期間)書き
  - Ex. C169(3Years),Months,Days,Hours

⑤ 4列目(IntervalLine)：　空
- ⑤にて() 書きする

⑥ 5列目(CodeOnly)：　1

⑦ 6列目(LineCoded)：　1

Ⅰ欄ア(Ia_c)：　0
Ⅰ欄イ(Ib_c)：　1
Ⅰ欄ウ(Ic_c)：　2
Ⅰ欄エ(Id_c)：　3
Ⅱ欄オ(II_c)：　5

20201226ver仕様書より
Ⅰ欄ア：　63 (64列目)
Ⅰ欄イ：　65 (66列目)
Ⅰ欄ウ：　67 (68列目)
Ⅰ欄エ：　69 (70列目)
Ⅱ欄オ：　71 (72列目)

1死亡票につき
1行～5行

行列番号は
0からカウント
しています