匿名加工情報の応用 (2):各種傷病を予測する健康診断モデル

池上 和輝[1,a)]　伊藤 聡志[1,b)]　菊池 浩明[2,c)]

概要：ビックデータの利活用が機械学習や AI 技術の発展により，企業・医療機関・金融機関など多様な場面で盛んになっている．また，個人情報を含んだデータにはプライバシー保護のための匿名加工技術が利用・研究されている．しかし，匿名加工データの有用性を評価する指標は定まっておらず，データの種類に依存する．そこで本稿では，あるヘルスケア企業が取得し，匿名加工した 20 万人の健康診断データと 29 万人のレセプトデータを突合し，ロジスティック回帰を行うことで健康診断データの 50 属性から傷病に関する統計的に有意な条件を抽出する．その条件における学習精度を評価し，既知の病気罹患に関する研究と比較することで有用性を確認する．また，250 種類の傷病について 3 年以内に罹患する確率を予測するモデルを健康診断データから作成する．最後に，健康診断データに対して追加の匿名加工を適用した場合の予測精度の変化を報告する．

キーワード：匿名加工，有用性評価，健康診断データ，レセプトデータ

Application of Anonymously Processed Information (2) : A Model Predicting Diseases given Medical Examination Data

Kazuki Ikegami[1,a)]　Satoshi Ito[1,b)]　Hiroaki Kikuchi[2,c)]

Abstract: Employing big data extensively has become popular in various scenes such as firms, medical institutions, and financial institutions due to the development of machine learning and AI technology. Technologies of de-identification for preventing individuals from being identified from the processed data have been studied. However, a utility loss incurred by de-identification is not clear. In this paper, we estimate the loss of utility via de-identified medical examination data and extract some significant factors from the health insurance claims associated with the medical records consisting of 200,000 and 290,000 individuals, respectively, by performing a logistic regression. Thereby the analysis results, we present some machine-learning models which predict the diseases that the given individual is likely to have in three years.

Keywords: De-identification, Utiliy, Medical Examination Data, Receipt Data

1. はじめに

ビックデータの利活用が機械学習や AI 技術の発展により，企業・医療機関・金融機関など多様な場面で盛んになっている．なかでも，健康診断データは病気の罹患を予測する有効な情報と考えられる．診断結果と傷病の関係を見るために，従来（個人情報保護法以前）は，人口動態統計死亡票の目的外利用を県の指導の元で承認してもらい，長期間のコホート研究を行なっていた．例えば，野田ら [5] は 10 万人について 8 年間追跡調査を行い，住民検診の検査結果とその後の脳卒中等による死亡の関係を明らかにした．また，日本人の健康寿命や生活習慣病に影響を与える要因を明らかにする目的で，国が全国で実施した循環器疾患基礎調査，および，国民健康・栄養調査の参加者を対象に追跡調査した NIPPON DATA[6] を用いたコホート研究が数多

[1] 明治大学大学院先端数理科学研究科
Nakano, Nakano-ku, Tokyo 164–8525, Japan
[2] 明治大学総合数理学部
Nakano, Nakano-ku, Tokyo 164–8525, Japan
[a)] cs192021@meiji.ac.jp
[b)] mmhm@meiji.ac.jp
[c)] kikn@meiji.ac.jp

く行われている．川南 [7] らは，喫煙習慣によるがん，肺がん死亡へ影響を分析し，非喫煙者に対する，毎日喫煙する集団の肺がん死亡の相対危険度が男性で 6.67 倍，女性で 3.67 倍であることを明らかにした．

しかし，2016 年の個人情報保護法の改正に伴い，個人情報は利用目的を特定して適切に取り扱うことが定められ，従来の様な死亡票の目的外利用によるコホート研究は困難になってきた．加えて，病歴などの情報は一般の個人情報よりも条件の厳しい要配慮個人情報に分類されることになり，個人の同意なく（オプトアウトなどでの）取得が禁止された．従って，大規模のコホートを長期に渡り追跡調査する従来の研究方法が出来なくなっている．

日本では 2017 年 5 月に改正個人情報保護法が施行され，要配慮個人情報を第三者に提供する際に，データに含まれる本人の同意をあらかじめとる（オプトイン）か，データの匿名加工が必要となった．匿名加工されたデータの評価は有用性と安全性の両面を考慮する必要があり，匿名加工データを評価する数多くの指標が提案されている．2017 年の同法施行後，数 100 の匿名加工情報取り扱い事業者が公表されているが，それを用いた活用例の報告は少ない．その理由には，加工の定量的な基準がない，健康診断データ特有の情報の取り扱いが難しいなどがある．

そでこ，本稿では，あるヘルスケア企業が実際に取得した 20 万人分の健康診断データと 28 万人分の傷病レセプトデータ（レセプトデータ），32 万人の適用データを使用する．これらのデータはいずれもヘルスケア企業によって匿名加工されている．

(目的) 匿名加工された健康診断データと傷病や生活習慣の相関を明らかにして，傷病罹患を予測モデルするモデルを作り，生活改善や健康施策作りに有益な知見を得ることである．

(方法)　健康診断とレセプトデータを突合し，ロジスティック回帰を行うことで健康診断データの 38 種類の特徴量から傷病に関する統計的に有意な条件を抽出する．また，傷病レセプト内の 274 種類の傷病について 4 種類の機械学習アルゴリズムを用いて，健康診断データから 3 年以内に各傷病に罹患する確率を予測するモデルを作成し，その精度を分析する．最後に，健康診断データに追加の k 匿名化を行い，追加加工データと元データとの予測精度を比較する．

(結果)　実験の結果，脳卒中の罹患には対象 38 説明変数のうち 22 説明変数が有意であった．例えば，十分に睡眠をとっている人は，脳卒中の罹患リスクが 0.89 倍に低下することを明らかにした．また，罹患予測モデルは傷病により精度のばらつきがあり，傷病間で F 値の差が最大 0.82 であった．k 匿名化によりサンプル数が 60%減少しても，F 値が高々 0.02 しか変化しないことを示した．

本研究の有用性と新規性を次に示す．

表 1　健康診断データの統計量

	対象年	レコード数	ユーザ数 N	欠損値セル数	特徴量数 M
処理前	2008-2018	964,635	198,740	10,536,861	49
処理後	2008-2016	203,521	68,629	0	38

(1) ロジスティック回帰分析により，34 説明変数と 274 の傷病の関係を明らかにしたこと．
(2) 4 種類の機械学習アルゴリズムを用いて，274 種類の傷病予測モデルを作成し，分類機よる精度を明らかにしたこと．傷病間の違いについても分析したこと．
(3) 健康診断データに追加の k 匿名を適用した時に，サンプル数が半分になっても罹患予測モデルの精度がほとんど劣化しないことを示したこと．

データの安全性については [1] で報告する．個人情報の取り扱いに関する倫理的配慮は 6 章で述べる．

2. データ

本研究では，あるヘルスケア企業が取得して匿名加工した健康診断データ，基本データ，傷病レセプトデータを使用する．健康診断データは，被験者 198,740 名の体重や身長等の身体的特徴 21 属性と問診結果 28 属性の計 49 属性の 10 年間分の健康診断結果から成る．傷病レセプトデータは，各患者が診断された傷病の記録である．また，基本データには被験者の生年月日や性別を示す．本研究では，基本データを用いて健康診断データに性別と年齢情報を付与したデータを健康診断データとして扱う．また，各データの詳細については [1] で報告する．

2.1　健康診断データ

健康診断データには，2008 年から 2018 年までの 20 万人分のデータが記録されている．分析の障害になる欠損値を含むレコードや相関が高い冗長な属性，カテゴリカル変数には次の前処理を行った．

(1) 欠損値レコードの多い 7 特徴量（列）を削除．
(2) 多重共線性 [2] をなくすために，相関係数が 0.7 以上ある 2 変数の一方を削除（4 特徴量）．
(3) 欠損値を含むレコード（行）の削除．
(4) カテゴリカル変数をダミー変数に変更．

処理前後の健康診断データの統計量を表 1 に，データに含まれる特徴量と削除したデータを図 1 に示す．図 1 のカッコ内の数字は上記の処理方法を表す．

本分析では，2.2 節のデータ処理のために 2008 年から 2016 年の健康診断データを使用し，各レコードを全て異なる被験者として分析する．また，健康診断データに含まれる被験者 id と受診日は削除する．

2.2　罹患情報の追加

本研究では，健康診断データとレセプトデータの相関を評価するために健康診断データ (健診）と罹患情報を突合

性別	年齢	身長	体重(2)	内臓脂肪面積(1)	bmi	腹囲実測(2)
収縮期血圧	拡張期血圧	中性脂肪	hdlコレステロール	ldlコレステロール	got ast	gpt alt(2)
ygtp	空腹時血糖(1)	hba1c ngsp	尿糖	尿蛋白	ヘマトクリット値(1)	血色素量(2)
赤血球数	クレアチニン(1)	尿酸(1)	健康分布(1)	メタボ判定	保健指導	服薬1血圧
服薬2血糖	服薬3脂質	既往歴1脳血管	既往歴2心血管	既往歴3腎不全等	貧血	喫煙
体重変化	運動習慣	身体活動	歩行速度	体重変化	食べ方1	食べ方2
食べ方3	食習慣	飲酒量(1)	飲酒	睡眠	生活習慣	保健指導の希望

実験で使用する連続値　実験で使用するカテゴリカル値

削除した特徴量

図 1　健康診断データの特徴量

する．

レセプトデータには，世界保健機関の医学分類リスト第 10 回修正国際疾患分類 (以下 ICD10)[3] に従った，傷病コードが記録されている．傷病コードは，ICD10 による大分類，中分類，細分類に分かれおり，例えば“血栓症による脳梗塞”の傷病は大分類コード I(循環器形)，中分類コード I63(I 脳梗塞)，細分類 I63.3(脳動脈の血栓症による脳梗塞) のように分類される．

レセプトデータに含まれる ICD10 の中分類コード (1490 種類) の傷病情報を次の条件で健康診断データに追加する．

$$\begin{cases} \text{健診の被験者 ID} = \text{レセプトの被験者 ID} \\ \text{健診受診年} \leq \text{傷病記録年} \leq \text{健診受診年} + 2 \end{cases} \tag{1}$$

被験者の罹患予測の情報活用として，健康的な食事や日常的な運動，十分な睡眠などの生活改善によりリスクを減らそうとすることが予想される．また，健康診断の結果と傷病の発病までには，一定の期間がかかると考えられる．そこで，生活習慣を改善し罹患を防止する期間を設け，発病までの期間を考慮するために 3 年の区間を定義した．また，本研究では，罹患者が 1,000 人以上の $D = 274$ 種類の傷病を分析対象とする．

表 2 に突合後のデータ例を示す．各傷病を目的変数 y，健康診断データを説明変数 x として分析を行う．被験者 i が傷病 A04 に罹患するかを健康診断データから予測するモデルは，

$$y_{A04} = f_{A04}(x_{i1}, x_{i2}, \ldots, x_{i38}) \tag{2}$$

で表される．ここで，f_{A04} は本分析で作成する機械学習モデルを表す．

罹患情報追加後の ICD10 についての罹患者数分布の上位 100 件を図 2 に示す．1490 種類の傷病のうち 16 % (225 種類) は罹患したレコードが 0 であった．付録 A.1 にレコード数上位 50 の傷病名等を示す．

3.　分析

3.1　分析方法

本稿では，3 つの分析を行う．1 つ目は，健康診断につ

表 2　データセット例

レコード ID	目的変数 (罹患有無)				説明変数 (健康診断結果)				
i	y_{A04}	y_{A09}	$\ldots$	y_{Z96}	x_1(年齢)	x_2(性別)	x_3(BMI)	$\ldots$	x_M
1	1	0		0	20	1	27.3		
2	0	0		0	31	0	23.1		
3	0	1		0	38	0	24.8		
⋮									
203,521	1	0		1	30	0	22.8		

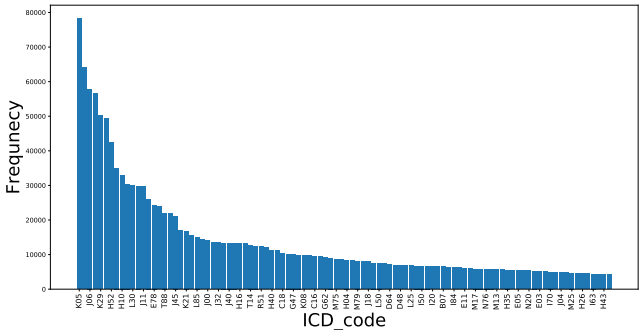

図 2　疾病ごとの罹患者数

いての傷病の罹患をロジスティック回帰する．統計的に有意となる診断結果を従来の研究と比較する．次に，3 年以内の罹患を予測する機械学習モデルを 274 種類の傷病について作成し，評価を行う．最後に，健康診断データに対し，k 匿名化を行い，加工による予測モデルの精度劣化を報告する．

3.2　健康診断と傷病の関係

傷病と因子の関係を明らかにした野田ら [5] が行った約 10 万人を対象とするコホート研究と比較する．彼らは 10 万人について 8 年間追跡調査を行い，住民健診の検査結果とその後の死亡の関係を男女別に COX 回帰分析を行い，統計的に有意な因子とその相対危険度を明らかにした．

我々の分析では，がん（ICD10 : C00-C99）と脳卒（ICD10 : I60-I69）を比較対象に使用し，3 年以内の罹患と説明変数（健康診断結果）の関係をロジスティック回帰を用いて分析する．

ある被験者 i の 3 年以内の傷病罹患確率 p_{iy} を

$$p_{iy} = \frac{1}{1 + e^{-z_i}} \tag{3}$$

で表す．ここで，z_i は健康診断データから得られる $M = 38$ 種類の説明変数 x と定数 α，各変数の係数 β について

$$z_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M \tag{4}$$

で定められる．

ある x_1 について，他の変数の影響を調整したオッズ比 (adjusted Odds Ratio) は，

$$OR = e^{\beta_1} \tag{5}$$

で与えられる．この分析では，罹患数が十分に小さい時，

表 3　先行研究との比較

	野田ら [5]	本分析
データ利用方法	人口動態統計死亡票の目的外使用	匿名加工情報
人数 N	92,277	68,629
説明変数数 M	12	37
傷病数 D	4	274
対象期間	1993–2001（9 年間）	2008–2016（9 年間）
被験者の年代	40 – 79	19 – 74
分析方法	cox 回帰	ロジスィック回帰
目的変数	死亡	三年以内の罹患

オッズ比と相対リスク（Rerative Risk）が等しいことを利用して，説明変数 x_1 による罹患影響をオッズ比 $OR = Pr($罹患 $|x_1 = 1)/Pr($非罹患 $|x_1 = 1)$ から確認する．

表 3 に野田らの実験と本分析の比較を示す．野田らの実験結果と比較する脳卒中とがんについては，母集団を先行研究と合わせるために健康診断データの 40 代以降のユーザーを抽出し，健康診断データの 38 特徴量，173,213 レコードを用いて分析を行う．また，他の傷病については母集団を全年代にするため 203,521 レコードを分析に使用する．分析には python の statsmodels ライブラリを用いる．

3.3　罹患予測モデル

健康診断データの有用性指標として，3 年以内の罹患予測モデルを傷病 $D = 274$ 種類作成する．学習時には罹患者数と同数の非罹患者レコードをランダムサンプリングして用いる．予測アルゴリズムには K 近傍法 (KNN)，RBF Support Vector Machine（SVM），Decision Tree (Tree)，Random Forest（RF）を使用する．各モデルの評価は 5 分割交差検証によって行い，有用性は再現率と適合率の調和平均である F 値の平均を使用する．モデルは python の scikit-learn を用いて実装し，ハイパーパラメータはデフォルト値を使用する．

3.4　匿名加工

匿名加工による 3.3 で提案した有用性指標の影響を明らかにする．本分析では，生活習慣に関する 13 特徴量を疑似識別子 (QI) として健康診断データを k 匿名化する．QI に使用した特徴量を表 4 に示す．表 4 中のメタボリックシンドロームの判定は，他の身体的特徴から診断されるため問診内容を省略している．$k = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$ で加工した時のデータに対して 3.3 節と同様の分析を行いモデルの精度を比較する．

4.　分析結果

4.1　ロジスティック回帰

表 A·2 にロジスティック回帰の結果を示す．estimate の正の値は罹患リスク増加，負の値は罹患リスク低下をそれぞれ表しており，*のついている値は統計的な有意差が確認できたものである．各 OR は，連続値の場合，値の増加による影響，2 値のカテゴリカル変数は 0 を基準に 1(質問に対して”はい”と答えた)，3 以上のカテゴリカル変数では最初の値をそれぞれ基準として各値のオッズ比を表している（estimate が 0.000 の値は，estimate が極めて小さい値と基準値を区別するための表記である）．例えば，脳卒中で睡眠の OR=0.890 から，睡眠が十分に取れている人 (睡眠 $x = 1$) は取れていない人 (睡眠 $x = 0$) に比べて脳卒中の 3 年以内罹患リスクが 0.89 倍である．3 年以内の脳卒中罹患リスクには 22 因子，がん罹患には 32 因子，インフルエンザには 25 因が有意であった．

また，表 A·2 の相対リスク RR は，野田ら [5] の研究結果を表す．ただし，BMI は 19 未満をベースとした時の 19 以上 21 未満の相対リスク，尿蛋白は+以上を尿蛋白異常とした時の尿蛋白正常 (–，±) に対する相対リスクである．

脳卒中の年齢，収縮期血圧では本分析の OR と野田らの RR から，同ほぼ等しい結果が得られていることがわかる．脳卒中とがんの両方で，既存研究と同様の結果が 5 項目から得られた．一方で，がんの喫煙による RR は既存研究が 1.51 に対して本分析では $OR = 0.964$ で不整合していた．この理由については考察で述べる．

また，先行研究には含まれなかった複数の問診結果（歩行又は身体活動，睡眠，食べ方 1，飲酒など）で有意な差が見られた．生活習慣の十分な睡眠をとることや 1 日 1 時間以上の歩行又は身体活動は，3 つ全ての疾病でリスクを下げる効果があった．

4.2　罹患推定モデルの評価

図 3 に 279 種類の疾病の罹患を予測した 4 種類のモデルの結果を示す．表 6 に各モデルの F 値の統計量を示す．統計量は，学習手法ごとに各 279 種類の傷病予測モデルの F 値を用いて計算される．ランダムフォレストの平均 F 値が最も高く 66%であった．一方で，他のモデルの平均 F 値は 57%で大きな差はなかった．SVM では標準偏差が他のモデルに比べて大きく 0.07 で，疾病により大きく精度が変化する．

図 3 から傷病の種類により精度の偏りがある．表 7 に分析で使用した中分類を大分類に再集計した統計量を示す．大分類の列は，対応する中分類の ICD コードを表し，中分類数の列は本分析で使用した中分類の傷病数を示している．また各学習手法の列は，中分類の平均 F 値を表す．表 7 から新生物，代謝疾患，尿路生起形疾患，妊娠，健康状態に影響をおよぼす要因等は他に比べて精度が高く，F 値が 0.7 である．

表 8 にランダムフォレストの F 値の上位 10 件の疾病を示す．サンプル数は，学習に使用したレコード数を表す．10 件中 9 件の傷病が女性特有の疾患であり，ランダムフォレスト以外のモデルでも精度が 70%以上だった．日本の老衰を除いた 3 大死亡原因 [4] であるがん，心疾患，脳血管

表 4　QI に使用した特徴量

説明変数	問診内容	解答形式
喫煙	現在、たばこを習慣的に吸っている	はい，いいえ
体重変化 20 歳からの	20 歳の時の体重から 10kg 以上増加している	はい，いいえ
運動習慣 30 分以上	1 回 30 分以上の軽く汗をかく運動を週 2 日以上、1 年以上実施	はい，いいえ
歩行又は身体活動	日常生活において歩行又は同等の身体活動を 1 日 1 時間以上実施	はい，いいえ
歩行速度	ほぼ同じ年齢の同性と比較して歩く速度が速い	はい，いいえ
体重変化 1 年間	この 1 年間で体重の増減が± 3 kg 以上あった	はい，いいえ
食べ方 2 就寝前	就寝前の 2 時間以内に夕食をとることが週に 3 回以上ある	はい，いいえ
食べ方 3 夜食間食	夕食後に間食（3 食以外の夜食）をとることが週に 3 回以上ある	はい，いいえ
食習慣	朝食を抜くことが週に 3 回以上ある	はい，いいえ
睡眠	睡眠で休養が十分とれている	はい，いいえ
飲酒	お酒（清酒、焼酎、ビール、洋酒など）を飲む頻度	ほとんど飲まない，時々，毎日
食べ方 1 早食い等	人と比較して食べる速度が速い	遅い，普通，早い
メタボリックシンドローム判定	-	非メタボ，メタボ予備軍，メタボ該当者

疾患 (脳卒中も含む) に該当する傷病の予測精度を表 9 に示す．脳梗塞は少なくとも 65%の精度で予測可能である．

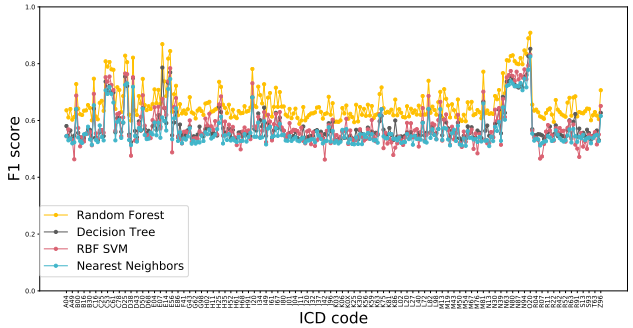

図 3　予測モデルの F 値

4.3　k 匿名化

表 10 に k の値によるレコード数の変化と，予測精度を示す．各学習手法の値は，279 種類の傷病を予測した際の F 値の平均を表す．10 から 100 の匿名化を行うと，最大で 60%のレコードが削除されても，予測精度が大きく変わらないことがわかる．279 種類の平均 F 値の絶対誤差は最大で 0.02 以下である．

5.　考察

5.1　コホート研究との比較

先行研究の RR と本分析の OR を比較した結果，6 因子中 5 因子が傷病リスクに対して同様に影響を示した．一方で，喫煙に関しては整合性が得られなかった．その原因として，日本人の喫煙率が先行研究の対象期間である 2001 年には 46%（男性）だったのに対し，本分析対象の 2016 年には 30%（男性）であった．このことから，健康診断で非喫煙と回答している人の中に元喫煙者が含まれていると予想される．この様な喫煙に関わる環境の変化の結果，喫煙による影響の整合性が保たれなかったと考える．

5.2　罹患予測モデル

表 8 の，一部の精度の高い疾病には女性特有（N97 女性不妊症，O20 妊娠早期の出血等）の傷病が多く罹患予測でなく男女推定の問題になっていた可能性がある．そこで図 4 に，全データと女性のみを使用したデータをそれぞれ学習させたランダムフォレストの結果を示す．散布図中のマーカーは各傷病の大分類を表す．N(尿路性器系の疾患) では，女性にのみのデータの時に精度が下がっている．従って，全体データを用いたモデルには男女推定の影響があったと考える．また，男女それぞれのデータからランダムフォレストで学習した結果を図 5 に示す．全体モデルから，男女別モデルでは精度劣化が確認されたが男女別のモデルでは精度劣化は確認されなかった．

図 3 の傷病によって精度が異なる原因には，I10（高血圧症）などの健康診断のと直接関係がある傷病と，T14(部位不明の損傷) など，健康診断とあまり関係がない傷病があるためと考える．

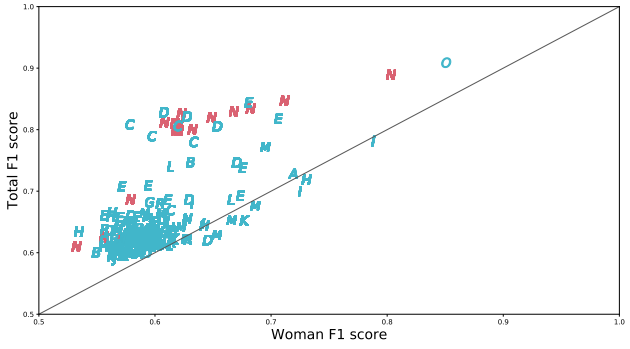

図 4　全データモデルと女性データモデルの精度比較

6.　個人情報の取り扱いに関する配慮

本研究では，健康診断データと疾病や生活習慣との相関を明らかにして疾病予防，生活改善，健康施策づくりに有

表 5 ロジスティック回帰結果

		脳卒中			がん			インフルエンザ	
	特徴量	estimate	OR	RR[5]	estimate	OR	RR[5]	estimate	OR
連続値	const	−2.980∗[1]	0.051		−1.515∗[1]	0.220		−1.957∗[1]	0.141
	年齢 (歳)	0.167∗[1]	1.182	1.14	0.119∗[1]	1.127	1.090	−0.255∗[1]	0.775
	身長 (cm)	−0.015	0.986		0.024∗[4]	1.025		0.024∗[4]	1.024
	Body Mass Index (kg/m^2)	−0.015	0.985	1.00	−0.054∗[1]	0.947	0.86	0.031∗[4]	1.031
	収縮期血圧 (mmHg)	0.033	1.033	1.02	−0.024∗[4]	0.976	–	−0.018	0.982
	拡張期血圧 (mmHg)	0.037∗[4]	1.037		−0.022∗[4]	0.979		−0.064∗[1]	0.938
	中性脂肪 (mg/dl)	0.041∗[2]	1.042		−0.025∗[3]	0.975		−0.005	0.995
	hdl コレステロール (mg/dl)	−0.003	0.997	‡	−0.021∗[3]	0.980	0.85	−0.006	0.994
	ldl コレステロール (mg/dl)	0.055∗[1]	1.057		−0.064∗[1]	0.938		−0.027∗[2]	0.973
	got ast (IU/L)	−0.011	0.989		0.038∗[1]	1.039		0.032∗[1]	1.032
	γ gtp (IU/L)	0.015	1.015		0.036∗[1]	1.036		0.001	1.001
	hba1c(ngsp)	0.042∗[3]	1.043		0.030∗[2]	1.030		0.044∗[1]	1.045
	赤血球数 ($*10^4/\mu l$)	−0.058∗[1]	0.943		−0.040∗[1]	0.961		0.008	1.009
2 値のカテゴリカル	性別	−0.102∗[1]	0.903		−0.256∗[1]	0.774		0.006	1.006
	服薬 1 血圧	0.127∗[1]	1.136	1.56	0.075∗[1]	1.078	1.15	0.042∗[1]	1.043
	服薬 2 血糖	0.045∗[1]	1.046		0.030∗[1]	1.030		−0.021∗[4]	0.979
	服薬 3 脂質	0.070∗[1]	1.073		0.034∗[1]	1.035		0.017∗[4]	1.017
	既往歴 1 脳血管	0.163∗[1]	1.177		−0.010	0.990		0.007	1.007
	既往歴 2 心血管	0.029∗[2]	1.029		−0.003	0.997		−0.011	0.989
	既往歴 3 腎不全・人工透析	0.007	1.007		0.028∗[1]	1.028		0.028∗[1]	1.029
	貧血	0.054∗[1]	1.056		0.071∗[1]	1.074		0.034∗[1]	1.034
	喫煙	0.008	1.008	1.27	−0.037∗[1]	**0.964**	**1.51**	0.027∗[2]	1.027
	体重変化 20 歳からの	0.025	1.025		0.005	1.005		0.034∗[2]	1.035
	運動習慣 30 分以上	−0.009	0.991		−0.025∗[2]	0.975		−0.007	0.993
	歩行又は身体活動	−0.044∗[2]	0.957		−0.052∗[1]	0.950		−0.036∗[1]	0.965
	歩行速度	−0.017	0.983		−0.050∗[1]	0.951		−0.018∗[4]	0.982
	体重変化 1 年間	0.055∗[1]	1.057		0.041∗[1]	1.042		0.047∗[1]	1.048
	食べ方 2 就寝前	0.035∗[3]	1.036		−0.003	0.997		0.032∗[1]	1.032
	食べ方 3 夜食間食	0.005	1.005		0.032∗[1]	1.033		0.017∗[4]	1.017
	食習慣	0.000	1.000		−0.037∗[1]	0.964		−0.025∗[2]	0.975
	睡眠	−0.117∗[1]	0.890		−0.057∗[1]	0.944		−0.087∗[1]	0.917
	保健指導の希望	0.036∗[3]	1.037		0.005	1.005		−0.005	0.995
3 値以上のカテゴリカル	メタボリックシンドローム判定								
	メタボ予備軍	0	1.000		0	1.000		0	1.000
	メタボ該当者	−0.038	0.963		−0.034∗[4]	0.967		−0.037∗[4]	0.964
	非メタボ	−0.033	0.968		−0.047∗[3]	0.954		−0.006	0.994
	食べ方 1（早食い等）								
	普通	0	1.000		0	1.000		0	1.000
	速い	0.044∗[1]	1.045		0.040∗[1]	1.041		0.012	1.012
	遅い	0.006	1.006		0.011	1.011		−0.009	0.991
	飲酒								
	時々	0	1.000		0	1.000		0	1.000
	ほとんど飲まない	0.039∗[3]	1.040		0.021∗[3]	1.021		0.009	1.009
	飲酒毎日	−0.015	0.985		−0.015	0.986		0.007	1.007
	保健指導レベル								
	情報提供	0	1.000		0	1.000		0	1.000
	動機付け支援	0.020	1.020		0.012	1.012		0.013	1.013
	対象外	−0.002	0.998		−0.034∗[4]	0.966		−0.058∗[3]	0.944
	積極的支援	0.030	1.030		0.015	1.015		−0.025	0.975
	尿糖								
	＋	0	1.000		0	1.000		0	1.000
	＋＋	−0.013	0.987		−0.003	0.997		−0.017	0.983
	＋＋＋	−0.027∗[4]	0.974		−0.009	0.991		−0.004	0.997
	-	−0.044∗[3]	0.957		−0.007	0.993		0.002	1.002
	±	−0.021	0.980		−0.003	0.997		0.005	1.005
	尿蛋白								
	＋	0	1.000	–	0	1.000	1.44	0	1.000
	＋＋	0.004	1.004		0.005	1.005		−0.016∗[4]	0.984
	＋＋＋	−0.010	0.990		−0.005	0.995		−0.007	0.993
	-	−0.034	0.966		−0.032∗[3]	0.968		−0.018	0.982
	±	0.007	1.007		0.010	1.010		0.003	1.003
	生活習慣の改善								
	改善予定 (1 か月以内)	0	1.000		0	1.000		0	1.000
	改善するつもりである (概ね 6 か月以内)	−0.047∗[3]	0.954		−0.012	0.988		0.010	1.010
	改善するつもりはない	−0.077∗[1]	0.926		−0.053∗[1]	0.948		−0.031∗[3]	0.970
	既に改善に取り組んでいる (6 ヶ月以上)	−0.019	0.981		−0.001	0.999		−0.008	0.992
	既に改善に取り組んでいる (6 ヶ月未満)	−0.009	0.991		−0.001	0.999		0.000	1.000

∗[1] : $P < 0.0001$, ∗[2] : $P < 0.001$, ∗[3] : $P < 0.01$, ∗[4] : $P < 0.05$

– : 有意な関連が全く示されたなかったため表示せず.

‡ : 分析モデルに含めなかったため表示せず.

表 6　各学習手法精度（F1）の統計量

	Mean	SD	Max	Min
RF	0.659	0.062	0.909	0.588
Tree	0.579	0.059	0.852	0.524
SVM	0.578	0.071	0.831	0.462
KNN	0.562	0.058	0.825	0.510

表 7　大分類での平均精度

大分類	傷病名	中分類数	RF	Tree	SVM	KNN
A00-B99	感染症および寄生虫症	15	0.642	0.563	0.557	0.551
C00-D48	新生物<腫瘍>	24	0.700	0.617	0.625	0.603
D50-D89	血液障害等	5	0.666	0.578	0.580	0.564
E00-E90	内分泌, 栄養および代謝疾患	15	0.711	0.624	0.628	0.595
F00-F99	精神および行動の障害	4	0.631	0.551	0.549	0.533
G00-G99	神経系の疾患	7	0.636	0.554	0.550	0.533
H00-H59	眼および付属器の疾患	16	0.652	0.570	0.589	0.552
H60-H95	耳および乳様突起の疾患	9	0.630	0.549	0.544	0.536
I00-I99	循環器系の疾患	18	0.673	0.587	0.589	0.562
J00-J99	呼吸器系の疾患	23	0.624	0.550	0.547	0.535
K00-K93	消化器系の疾患	34	0.631	0.554	0.547	0.543
L00-L99	皮膚および皮下組織の疾患	20	0.638	0.558	0.563	0.544
M00-M99	筋骨格系および結合組織の疾患	24	0.645	0.565	0.562	0.550
N00-N99	尿路性器系の疾患	23	0.746	0.669	0.673	0.648
O00-O99	妊娠, 分娩および産じょく	1	0.909	0.852	0.831	0.825
Q00-Q99	先天奇形, 変形および染色体異常	1	0.656	0.569	0.564	0.554
R00-R99	異常検査所見で他に分類されないもの	24	0.634	0.559	0.541	0.537
S00-T98	損傷, 中毒およびその他の外因の影響	10	0.624	0.549	0.535	0.538
Z00-Z99	健康状態に影響をおよぼす要因等	1	0.707	0.627	0.651	0.616

表 8　F 値上位 10 件の疾病

ICD10	傷病名	サンプル数	RF	Tree	SVM	KNN
O20	妊娠早期の出血	2,844	0.909	0.852	0.831	0.825
N97	女性不妊症	2,374	0.889	0.826	0.794	0.778
E10	1 型糖尿病	2,000	0.869	0.786	0.676	0.611
N94	月経周期の疼痛等	3,322	0.847	0.753	0.780	0.747
E28	卵巣機能障害	11,204	0.844	0.770	0.784	0.746
N95	閉経期障害等	6,564	0.835	0.760	0.745	0.717
N80	子宮内膜症	4,066	0.830	0.746	0.757	0.730
D25	子宮平滑筋腫	14,814	0.828	0.755	0.765	0.725
N76	膣及び外陰のその他の炎症	11,608	0.827	0.757	0.774	0.738

表 9　日本 3 大死亡原因の罹患予測精度

ICD10	傷病名	サンプル数	RF	Tree	SVM	KNN
C18	結腸がん	20,470	0.604	0.531	0.538	0.524
I20	狭心症	13,178	0.652	0.570	0.580	0.543
I63	脳梗塞	8,806	0.648	0.565	0.587	0.545

表 10　健康診断データの統計量

k	レコード数	削除割合	RF	Tree	SVM	KNN
0	203,521	0.00	0.659	0.579	0.578	0.562
10	167,682	0.18	0.654	0.577	0.574	0.561
20	145,918	0.28	0.652	0.575	0.572	0.560
30	130,535	0.36	0.651	0.576	0.570	0.560
40	118,668	0.42	0.648	0.573	0.564	0.558
50	110,592	0.46	0.647	0.574	0.564	0.558
60	103,301	0.49	0.645	0.572	0.561	0.556
70	97,696	0.52	0.644	0.573	0.560	0.555
80	92,878	0.54	0.645	0.573	0.559	0.556
90	87,920	0.57	0.642	0.570	0.555	0.556
100	84,120	0.59	0.642	0.571	0.559	0.556

益な知見を得ることを目的に，匿名加工情報（個人情報の保護に関する法律（平成 15 年法律第 57 合）第 2 条 9 項）を用いている．同法，関連する法令，ガイドラインなどを遵守して，適切な安全管理措置を施して研究を遂行している．本稿で発表する研究結果には，特定の個人を識別可能な情報が含まれず，健康診断被験者のプライバシーへ及ぼす影響がないことを，事前に (2020 年 7 月 30 日) ヘルスケア企業に相談，確認済みである．本匿名加工情報は第三者

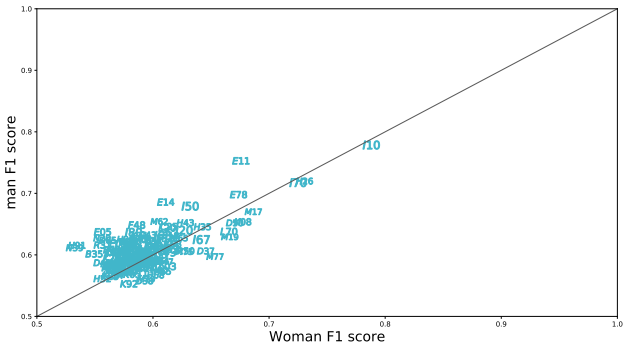

図 5　男女別データモデルの精度比較

提供する予定はない．ガイドライン [8] 第 12 の 2「研究の成果の公表にあたっての留意点」に抵触している該当項目はないことを確認している．

7. おわりに

本稿では，あるヘルスケア企業から提供された匿名加工健康診断データの有用性を評価した．先行研究で示された傷病と因子の結果を比較して，匿名加工情報はコホート研究の代わりとなる有用なデータであることを確認した．本分析により，脳梗塞の三年以内の罹患リスクが飲酒をほとんどしない人は，時々飲酒をする人に比べて 1.04 倍高くなることや，十分な睡眠を取ることでリスクを 0.89 倍に下げるなどの新たな知見を得た．また，健康診断データと傷病レセプトデータから 279 種類の傷病に 3 年以内に罹患するモデルをそれぞれ 4 種類の機械学習手法を用いて作成した．ランダムフォレストが最も予測精度が良く 279 種類の傷病の平均 F 値は 0.65 であった．さらに，生活習慣に関する 13 特徴量を QI として K=10 から 100 の追加の匿名化を行い予測モデルの精度の変化を確認した．K=100 の時レコード数は最大で 60%減少するが，F 値は最大で 0.02 しか変化せず，加工しても十分に精度良いモデルが作れることを示した．

罹患予測モデルの特徴量に被験者の過去の健康診断データを使用すること，学習モデルの最適なパラメーターを探索することを今後の課題とする．

謝辞

匿名加工情報を提供頂いたヘルスケア企業に感謝する．本研究の倫理面の助言を頂いた CSS 研究倫理相談 TF に感謝する．

参考文献

[1] 伊藤 聡志，池上 和輝，菊池 浩明匿名加工情報の応用 (1): 健康診断データとレセプトデータ の分析とプライバシーリスク評価，(CSS2020 にて発表予定)．
[2] 松井秀俊，小泉和之，統計モデルと推測，講談社，2019，

p103.
[3] 厚生労働省：疾病、傷害及び死因の統計分類 (https://www.mhlw.go.jp/toukei/sippei/，2020.08.14 参照)
[4] 厚生労働省：平成 29 年人口動態統計月報年計の概況 (https://www.mhlw.go.jp/toukei/saikin/hw/jinkou/geppo/nengai17/index.html，2020.08.14 参照)
[5] 野田 博之，磯 博康，西連地利己，入江ふじこ，深澤 伸子，鳥山 佳則，大田 仁史，能勢 忠男，住民健診 (基本健康診査) の結果に基づいた脳卒中・ 虚血性心疾患・全循環器疾患・がん・総死亡の予測，日本公衛誌 53: 265–277，2006.
[6] NIPPON DATA (https://shiga-publichealth.jp/nippon-data/，2020.08.14 参照)
[7] 川南 勝彦，簑輪 眞澄，岡山 明、早川 岳人，上島 弘嗣，NIPPON DATA80 研究グループ，喫煙習慣の全死因，がん，肺がん死亡への影響に関する研究：NIPPON DATA80，日本衛生学雑誌.2003 Jan;57(4):669-673.
[8] 厚生労働省，レセプト情報・特定健診等情報の提供に関するガイドライン，平成 23 年（平成 28 年改訂）

付　　録

A.1　主要な傷病一覧

表 A·1　罹患者数上位 50 の傷病名とモデル精度

ICD10	傷病名	レコード数	RF	Tree	SVM	KNN
K05	歯肉炎及び歯周疾患	156,394	0.600	0.531	0.543	0.525
K02	う＜齲＞蝕	128,304	0.595	0.529	0.534	0.521
J06	多部位及び部位不明の急性上気道感染症	115,698	0.609	0.538	0.563	0.529
J30	血管運動性鼻炎	113,320	0.619	0.544	0.565	0.530
K29	胃炎及び十二指腸炎	100,346	0.597	0.533	0.540	0.517
J20	急性気管支炎	98,660	0.611	0.541	0.562	0.529
H52	屈折及び調節の障害	84,688	0.610	0.540	0.560	0.529
J02	急性咽頭炎	70,022	0.608	0.538	0.560	0.528
H10	結膜炎	66,128	0.631	0.551	0.579	0.536
K04	歯髄及び根尖部歯周組織の疾患	60,868	0.595	0.536	0.529	0.519
L30	その他の皮膚炎	59,804	0.621	0.545	0.569	0.535
A09	その他の胃腸炎及び大腸炎	59,604	0.611	0.543	0.569	0.530
J11	インフルエンザ	59,402	0.600	0.537	0.575	0.535
E14	詳細不明の糖尿病	52,226	0.686	0.591	0.620	0.565
E78	リポタンパク代謝障害	48,748	0.692	0.593	0.648	0.608
K76	その他の肝疾患	47,782	0.626	0.546	0.557	0.537
T88	外科的及び内科的ケアの合併症	44,070	0.605	0.531	0.551	0.527
M54	背部痛	43,848	0.597	0.536	0.525	0.511
J45	喘息	42,234	0.625	0.549	0.564	0.528
J01	急性副鼻腔炎	33,812	0.641	0.556	0.591	0.545
K21	胃食道逆流症	33,622	0.612	0.537	0.534	0.517
I10	本態性高血圧（症）	31,172	0.782	0.681	0.731	0.681
L85	その他の表皮肥厚	30,180	0.637	0.550	0.585	0.549
J10	その他のインフルエンザウイルスが分離されたインフルエンザ	28,656	0.588	0.530	0.553	0.527
J00	急性鼻咽頭炎［かぜ］＜感冒＞	28,110	0.614	0.543	0.552	0.529
J03	急性扁桃炎	27,068	0.625	0.545	0.567	0.535
J32	慢性副鼻腔炎	27,024	0.625	0.546	0.558	0.534
A49	部位不明の細菌感染症	26,730	0.603	0.534	0.532	0.519
J40	気管支炎	26,706	0.607	0.538	0.536	0.519
H16	角膜炎	26,596	0.639	0.552	0.598	0.550
K25	胃潰瘍	26,414	0.594	0.530	0.529	0.516
T14	部位不明の損傷	25,506	0.601	0.532	0.545	0.520
M47	脊椎症	24,932	0.612	0.538	0.555	0.528
R51	頭痛	24,562	0.617	0.544	0.564	0.526
K03	歯の硬組織のその他の疾患	24,084	0.596	0.531	0.549	0.528
H40	緑内障	22,444	0.626	0.552	0.551	0.531
B35	皮膚糸状菌症	22,230	0.601	0.539	0.513	0.514
C18	結腸の悪性新生物＜腫瘍＞	20,470	0.604	0.531	0.538	0.524
K59	その他の腸の機能障害	20,332	0.651	0.561	0.601	0.552
G47	睡眠障害	20,262	0.624	0.544	0.537	0.528
M51	その他の椎間板障害	19,754	0.598	0.531	0.507	0.512
K08	歯及び歯の支持組織のその他の障害	19,612	0.653	0.566	0.606	0.570
D50	鉄欠乏性貧血	19,526	0.746	0.644	0.689	0.641
C16	胃の悪性新生物＜腫瘍＞	18,960	0.614	0.541	0.555	0.532
E86	体液量減少（症）	18,820	0.619	0.541	0.544	0.528
G62	その他の多発（性）ニューロパチ＜シ＞ー	18,302	0.607	0.542	0.538	0.519
R11	悪心及び嘔吐	18,032	0.638	0.558	0.584	0.545
M75	肩の傷害＜損傷＞	17,034	0.623	0.539	0.554	0.535
K12	口内炎及び関連病変	17,010	0.624	0.547	0.565	0.534
H04	涙器の障害	16,816	0.666	0.574	0.624	0.568