情報処理学会論文誌 Vol.62 No.9 1560–1574 (Sep. 2021)

健康診断データとレセプトデータの匿名加工情報を用いた疾病リスク分析

伊藤 聡志[1,2,a)] 池上 和輝[1,b)] 菊池 浩明[3,c)]

受付日 2020年11月18日, 採録日 2021年6月7日

概要：健康診断は非常に有用なデータであり，個人がこれから罹患する危険性のある病気を予測できる可能性がある．近年の法制度の改正やプライバシ意識の高まりを受けて，個人情報保護法改正によって利用目的を明確にしないで個人データを取得することを禁じられたため，診断結果と傷病の関係を見る主流の手法であった長期間のコホート研究が困難になった．そこで，この問題を解決する手法として，本研究では匿名加工情報に着目する．本稿では，あるヘルスケア企業が取得した 10 年間の 20 万人分の健康診断データと 28 万人分のレセプトデータのからなる匿名加工情報を用いて，従来のコホート研究に類するような新しい分析が可能であるかどうかを研究する．相対リスクやロジスティック回帰を用いて疾病の罹患リスクの分析を行い，複数の予測アルゴリズムを用いて 3 年以内の罹患を予測するモデルを作成する．また，健康診断データをいくつかの疑似識別子について k–匿名化することにより，分析精度がどれだけ変化するのかを明らかにする．

キーワード：匿名加工，再識別，健康診断データ，レセプトデータ，相対リスク，オッズ比，k–匿名

Analysis of Diseases Risk for Anonymously Processed Information of Medical Examination and Health Insurance Claims

Satoshi Ito[1,2,a)] Kazuki Ikegami[1,b)] Hiroaki Kikuchi[3,c)]

Received: November 18, 2020, Accepted: June 7, 2021

Abstract: A medical examination is very useful and helps predicting diseases risks for patients. Long-term cohort studies have been made to predict diseases risk. However, the cohort study becomes difficult since the Act on the Protection of Personal Information fully came into effect in Japan, where it is prohibited to obtain personal data without specifying clear purpose of use. In this paper, we focus on the Anonymously Processed Information to address this problem. We analyze an anonymously processed information to predict the diseases risk using the anonymously processed information of medical examination and health insurance claim data consisting of 200,000 and 290,000 individuals. We aim to prove that an Anonymously Processed Information would be a new methodology as well as the conventional cohort data. We analyze a diseases risk by performing a logistic regression and calculating a relative risk and develop some machine-learning models which predict the likelihood of the diseases in three years given individual medical examination. Finally, we reveal how much the accuracy of analysis is reduced by k–anonymizing data.

Keywords: de-identification, re-identification, medical examination data, receipt data, relative risk, odds ratio, k–anonymity

1. はじめに

機械学習や AI 技術の発展により，ビックデータの利活用が企業・医療機関・金融機関等，多様な場面でさかんに

1 明治大学大学院先端数理科学研究科
Graduate School of Advanced Mathematical Sciences, Meiji University, Nakano, Tokyo 164–8525, Japan
2 日本学術振興会
Japan Society for the Promotion of Science, Chiyoda, Tokyo 102–0083, Japan
3 明治大学総合数理学部
School of Interdisciplinary Mathematical Sciences, Meiji University, Nakano, Tokyo 164–8525, Japan

a) mmhm@meiji.ac.jp
b) cs192021@meiji.ac.jp
c) kikn@meiji.ac.jp

なっている．なかでも健康診断は非常に有用であり，過去の統計に基づいて個人がこれから罹患する危険性のある病気を予測できる可能性がある．診断結果と傷病の関係を見るために，2016 年の個人情報保護法の改正前は長期間のコホート研究が主流であった．たとえば，野田らは厚生労働省と総務省の許可を得て人口動態統計死亡票を目的外利用して，茨城県に住む 92,277 人の住民健診データを分析することにより，検査項目と死亡との関係を相対リスク（relative risk）等を用いて明らかにした [1]．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 [2]，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 [3] 等の大規模コホート研究が数多く行われていた．川南ら [4] 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リスクが男性で 6.67 倍，女性で 3.67 倍であることを明らかにした．

しかしながら，近年の法制度の改正やプライバシ意識の高まりを受けて，個人情報保護法 [27] では，利用目的を明確にしないで個人についてのデータを取得することを禁じており，特に検査結果や病歴等は要配慮個人情報に分類され，特別な措置を必要とされている[*1]．そのため，従来のよう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．

そこで，この問題を解決する手法として，本研究では匿名加工情報に着目する．個人が識別されることを防ぐために個人情報を加工する技術を匿名化という．個人情報保護法では，法律施行規則 19 条 1–5 号までの要件を満たして加工された情報を匿名加工情報という．匿名加工情報から個人を識別しようとする再識別は法で禁じられている．個人情報保護法では，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，個人情報の第三者提供とならないようにデータを匿名加工情報とすることが必要となった．

匿名加工情報は，従来のコホート研究よりも手軽に大規模のデータを集めて様々な分析を試行できる利点がある．その反面，匿名加工を行う際には，特異な記述等の削除や他の個人情報データベースの性質を勘案して適切な処置を講じること（規則 19 条）が求められており，これらの加工が分析結果に影響を及ぼすことが懸念される．現に，匿名加工情報を生成していることを公表している取扱事業者は 300 社を超えている [5] が，加工方法の多様性や加工の度合いの自由度があり，実際に利活用している企業は 25%にすぎないという報告 [6], [7] がある．

そこで本稿では，あるヘルスケア企業が取得した 10 年間の 20 万人分の健康診断データと 28 万人分のレセプトデータのからなる匿名加工情報を用いて，従来のコホート研究と並ぶ有用な分析が可能であるかどうかを研究する．健康診断データには，各個人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 属性の健康診断結果が記録されている．一方，レセプトデータには，各個人に処方された医薬品の情報が記録された医薬品レセプトデータ（21 属性）と，各個人が診断された傷病の情報が記録された傷病レセプトデータ（15 属性）の 2 種類がある．これらのデータはいずれも当該ヘルスケア企業によって法律に従って適切に匿名加工されたものであるが，これらを分析することによって，様々な傷病の罹患リスク等の有用な知見を得ることが期待できる．当該ヘルスケア企業による匿名加工情報は法律上は十分であるが，将来，再識別攻撃が新たに発見され，個人識別の脅威が問題になるときのために，さらに識別性を低減させる加工を重ねて，有用性の低下を評価しておくことが望ましい．そこで，本稿ではこれらの匿名加工情報にさらなる加工を施した後に分析をして，匿名加工が分析結果へ及ぼす影響を明らかにする．

本研究の目的は次のとおりである．(1) 健康診断データと傷病や生活習慣の相関を明らかにし，傷病罹患を予測するモデルを作り，生活改善や健康施策作りに有益な知見を得ること．(2) 匿名加工情報を用いても，ヘルスケア情報の分析結果として有用性が認められるレベルの品質の結果が得られるかどうかを検証すること．

このために，次の方法により匿名加工情報の分析を行う．

(1) 健康診断データと，診断されたことのある傷病/処方されたことのある傷病/医薬品のレセプトデータをクロス集計して，疾病の相対リスクを分析する．
(2) がんと脳卒中を対象とし，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して，従来のコホート研究結果と比較する．
(3) K 近傍法（KNN），RBF Support Vector Machine（SVM），Decision Tree（Tree），Random Forest（RF）を予測アルゴリズムに使い，3 年以内の罹患を予測するモデルを 274 種類の傷病について作成する．
(4) 健康診断データをいくつかの疑似識別子（QI: Quasi Identifier）について k–匿名性（k–anonymity）を満たす複数のアルゴリズムによる加工を行い，分析精度がどれだけ変化するのかを明らかにする．

上記の分析から，導かれた本研究の主要な結論は以下のとおりである．

(1) 高血圧を危険因子としたときの循環器系の疾患の相対リスクが 1.75 であることを明らかにした．1 度もレセプトに記録がない（病院にかかっていない）特異な集団「健康集団」があるが，その健康診断結果はそれほど健康ではない．
(2) 十分な睡眠をとる人が 3 年以内に脳卒中となる罹患リ

[*1] ただし，個人情報の保護に関する法律 [27] では学術研究，行政機関の保有する個人情報の保護に関する法律 [28] では相当な理由がある，または業務遂行に必要な限度，あるいは学術研究での利用や提供が可能となっている．

表 1 先行研究との比較
Table 1 The comparison with previous study.

	野田ら [1]	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	ロジスティック回帰 機械学習等
目的変数	死亡	3 年以内の罹患

スクが，睡眠不足の人に比べて 0.787 倍になることや，加齢による脳卒中のリスクが文献 [1] と整合した結果が得られたこと．

(3) ランダムフォレストが最も予測精度が良く，274 種類の傷病の平均 F 値は 0.65 である．

(4) 性別・年齢を QI として $k = 1{,}000$ までの追加の k–匿名化をした結果，$k = 1{,}000$ のときレコード数は約 10%減少するが，加工しても 274 種類の最大誤差は 0.007 であり，十分に精度良いモデルが作れることを示した．病歴を QI とした k–匿名化においても，相対リスクの相対誤差が $k = 10$ で 0.073 であり，十分な精度を保持する．

(5) 匿名加工情報と当該ヘルスケア企業より確認した予測健診データの OR の誤差は，平均 $2.5 \cdot 10^{-4}$ であった．

傷病と因子の関係を明らかにした野田ら [1] が行った約 10 万人を対象とするコホート研究と本研究の比較を，**表 1** に示す．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 比例ハザードモデルを用いて偏回帰係数を求める分析を行い，統計的に有意な因子とその相対危険度を明らかにした．一方，本研究は匿名加工情報を活用することで従来の 4 種から 274 種の多くの疾病について分析することが可能になった．

本稿では，2 章で健康診断データと傷病/医薬品レセプトデータについての説明を行い，3 章でこれらのデータの分析を行う．

2. 健康診断データと傷病/医薬品レセプトデータ

2.1 概要

本稿で分析する健康診断データの個人・レコード・属性数を**表 2** に示し，各属性を**表 3** に示す．第 3–17，20–24 属性には連続値が，それ以外の属性には離散値が記録されている．第 1–27 属性は個人の身体情報を示し，第 28 属性以降は個人の問診 28 問 [9] への回答結果を示している．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．

第 25 属性の「健診ランク」は，bmi や中性脂肪等の 12 属性から個人のリスクを判定する指標の分布であり，A（非肥満）と B（肥満），1（リスクなし）–4（服薬投与）を組み合わせた 8 ランクに分類される．健康診断データにおける健診ランク（レコード）を**表 4** に示す．13.2%のレコードが最も健康なランクである A1（非肥満・リスクなし）に，9.3%のレコードが最も不健康なランク B4（肥満・服薬投与）に該当した．また，この属性の情報を持たないレコードも多く，全体の 43.8%が “不明” となっていた．

表 2 3 データの統計情報
Table 2 The statistics of three data.

データ名	健康診断データ	傷病レセプト	医薬品レセプト
個人数 n	198,740	288,568	279,199
レコード数	964,636	39,363,878	31,465,504
属性数	49	15	21
レセプト枚数	–	11,912,236	9,000,249
対象年	2008–2018	2012–2018	2012–2018

2.2 レセプトデータ

本レセプトデータには，各個人が診断された傷病の詳細が記録されている傷病レセプトデータと，各個人が処方された医薬品の詳細が記録されている医薬品レセプトデータの 2 種類がある．傷病/医薬品レセプトデータの統計量を表 2 に示す．

傷病レセプトデータの第 7–12 属性と医薬品レセプトデータの第 14–17 属性は傷病/医薬品分類コードである．傷病の分類コードには国際疾病分類第 10 版（ICD10）[10] が，医薬品の分類コードには解剖治療化学分類（ATC 分類）[11] が用いられており，これらの分類コードは大分類 > 中分類 > 小分類 > 細分類とカテゴリ分けされている．たとえば脳梗塞という病気は，循環器系の疾患（大分類コード：*I*）の中の脳梗塞カテゴリ（中分類コード：*I63*）の中の脳梗塞（細分類コード：*I639*）に分類される．

各レセプトデータは複数のレコードからなる．表 2 から，傷病レセプトでは平均 3.3 レコード/枚，医薬品レセプトでは平均 3.50 レコード/枚がある．しかし，**図 1** に傷病レセプトデータにおける各顧客ごとのレコード数分布（降順）を示すように，一様ではない．上位 9 人の個人のレコード数が飛びぬけて多く，歪んでいる．10 位の個人のレコード数が 4,015 であるのに対し，9 位の個人のレコード数は 321,828 であり，1 位の個人は 2,588,244 レコードも記録されている．

レセプトの枚数についても同様に歪んでおり，1 位の個人は 1 人で 855,147 枚のレセプトを処方されている．上位 9 人の頻度とレセプト数が飛びぬけて多いことは，医薬品レセプトデータにおいても同じことがいえる[*2]．

[*2] 仮個人 id と仮レセプト id について分析を行った結果，1 枚のレセプトが 2 人の個人に対応するケースが存在した（傷病：8,275 枚，医薬品：6,504 枚）．仮レセプト id 属性は仮名化されたものであるため，その際に重複が生じた可能性がある．

表 3 健康診断データに記録されている情報
Table 3 The details of 49 attributes of the medical examination data.

index	種類	属性名	欠損値数	一意な値の数	平均識別確率
1	離散/身体	仮個人 id	0	–	–
2	離散/身体	健診受診月	0	1	$1.31 \cdot 10^{-4}$
3	連続/身体	身長	1,048	5	$7.75 \cdot 10^{-4}$
4**	連続/身体	体重	1,060	19	$1.14 \cdot 10^{-3}$
5*	連続/身体	内臓脂肪面積	964,296	262	$3.15 \cdot 10^{-4}$
6	連続/身体	bmi	1,065	0	$3.60 \cdot 10^{-4}$
7**	連続/身体	腹囲 実測	76,519	28	$8.46 \cdot 10^{-4}$
8	連続/身体	収縮期血圧	154,021	0	$1.43 \cdot 10^{-4}$
9	連続/身体	拡張期血圧	154,023	0	$1.04 \cdot 10^{-4}$
10	連続/身体	中性脂肪	29,740	192	$1.38 \cdot 10^{-3}$
11	連続/身体	hdl コレステロール	29,765	173	$4.11 \cdot 10^{-4}$
12	連続/身体	ldl コレステロール	29,922	116	$4.00 \cdot 10^{-4}$
13	連続/身体	got ast	28,140	7	$2.01 \cdot 10^{-4}$
14**	連続/身体	gpt alt	28,141	5	$2.56 \cdot 10^{-4}$
15	連続/身体	γ gtp	28,161	88	$8.13 \cdot 10^{-4}$
16*	連続/身体	空腹時血糖	372,933	5	$2.81 \cdot 10^{-4}$
17	連続/身体	hba1c ngsp	111,921	15	$1.22 \cdot 10^{-4}$
18	離散/身体	尿糖	6,177	0	$1.21 \cdot 10^{-5}$
19	離散/身体	尿蛋白	5,301	0	$1.21 \cdot 10^{-5}$
20*	連続/身体	ヘマトクリット値	444,694	0	$3.72 \cdot 10^{-4}$
21**	連続/身体	血色素量	332,031	0	$1.54 \cdot 10^{-4}$
22	連続/身体	赤血球数	331,553	2	$3.74 \cdot 10^{-4}$
23*	連続/身体	クレアチニン	746,905	150	$3.83 \cdot 10^{-4}$
24*	連続/身体	尿酸	741,879	2	$1.25 \cdot 10^{-4}$
25*	離散/身体	健診ランク	422,239	0	$2.34 \cdot 10^{-5}$
26	離散/身体	メタボリックシンドローム判定	143,700	0	$1.18 \cdot 10^{-5}$
27	離散/身体	保健指導レベル	154,261	0	$1.40 \cdot 10^{-5}$
28	離散/問診	服薬 1 血圧	58,424	0	$1.18 \cdot 10^{-5}$
29	離散/問診	服薬 2 血糖	58,512	0	$1.16 \cdot 10^{-5}$
30	離散/問診	服薬 3 脂質	58,520	0	$1.13 \cdot 10^{-5}$
31	離散/問診	既住歴 1 脳血管	350,483	0	$1.20 \cdot 10^{-5}$
32	離散/問診	既住歴 2 心血管	350,393	0	$1.19 \cdot 10^{-5}$
33	離散/問診	既住歴 3 腎不全・人工透析	350,590	0	$1.14 \cdot 10^{-5}$
34	離散/問診	貧血	351,960	0	$1.18 \cdot 10^{-5}$
35	離散/問診	喫煙	40,513	0	$1.16 \cdot 10^{-5}$
36	離散/問診	体重変化 20 歳からの	356,876	0	$1.21 \cdot 10^{-5}$
37	離散/問診	運動習慣 30 分以上	205,592	0	$1.01 \cdot 10^{-5}$
38	離散/問診	歩行または身体活動	205,783	0	$9.73 \cdot 10^{-6}$
39	離散/問診	歩行速度	357,278	0	$1.16 \cdot 10^{-5}$
40	離散/問診	体重変化 1 年間	371,610	0	$1.01 \cdot 10^{-5}$
41	離散/問診	食べ方 1 早食い等	357,880	0	$1.43 \cdot 10^{-5}$
42	離散/問診	食べ方 2 就寝前	205,902	0	$1.01 \cdot 10^{-5}$
43	離散/問診	食べ方 3 夜食・間食	220,089	0	$9.62 \cdot 10^{-6}$
44	離散/問診	食習慣	207,343	0	$1.06 \cdot 10^{-5}$
45	離散/問診	飲酒	271,688	0	$1.44 \cdot 10^{-5}$
46*	離散/問診	飲酒量	459,731	0	$1.52 \cdot 10^{-5}$
47	離散/問診	睡眠	357,548	0	$1.11 \cdot 10^{-5}$
48	離散/問診	生活習慣の改善	364,315	0	$1.54 \cdot 10^{-5}$
49	離散/問診	保健指導の希望	356,536	0	$1.13 \cdot 10^{-5}$

*：2.4 節のクレンジング手法 1 で削除を行った．
**：2.4 節のクレンジング手法 2 で削除を行った．

2.3 傷病/医薬品と健康診断データ

3 つのデータ（健康診断データ，傷病レセプトデータ，医薬品レセプトデータ）を用いることにより，個人を診断されたことのある傷病/処方されたことのある傷病/医薬品と健康診断データをクロス集計する．3 つのデータは，同一の個人情報取扱事業者により加工された単一の匿名加工情報である．法令に従った規則性を有しない方法で生成された，共通の仮 ID が振られている．**図 2** に 3 データ間の包含関係をベン図で示す．3 データすべてに記録されている個人は 178,033 人であり，傷病/医薬品レセプトデータにしか記録されていない個人も存在した．傷病/医薬品グループ間には個人の重複があり，個人は複数のグループに属することができる．

図 3 に傷病グループごとの健診ランクを示す．x 軸は傷病分類コード（大分類）を意味しており，“*He*” は 3.2.3 項

表 4 健康診断データにおける健診ランクの分布
Table 4 The health degree distribution in the medical examination data.

状態	健診ランク	レコード数	割合
非肥満	A1	127,550	0.132
	A2	87,487	0.091
	A3	45,155	0.047
	A4	66,744	0.069
肥満	B1	24,573	0.025
	B2	48,367	0.050
	B3	52,726	0.055
	B4	89,794	0.093
不明	不明	422,239	0.438

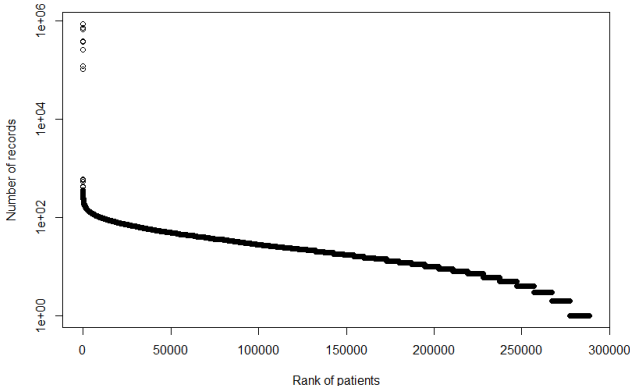

図 1 傷病レセプトデータにおける各顧客ごとのレコード数分布
Fig. 1 The distribution of the number of records in disease health insurance claim data.

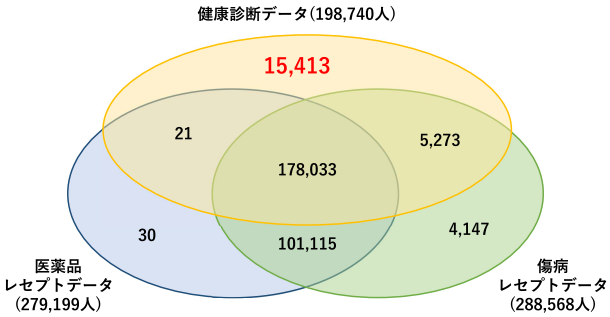

図 2 3 データ間の包含関係
Fig. 2 The relation among three data.

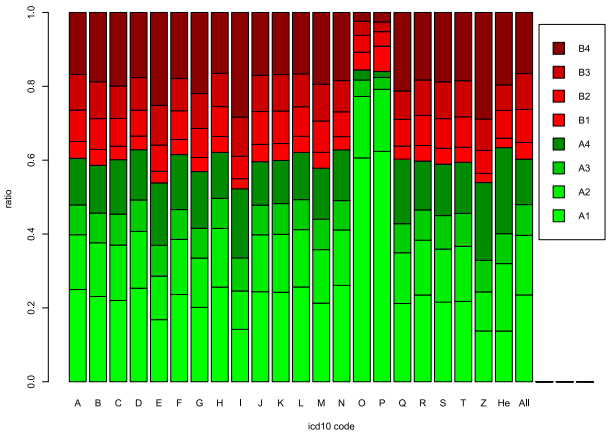

図 3 傷病グループごとの健診ランク
Fig. 3 The health degree distribution for disease groups.

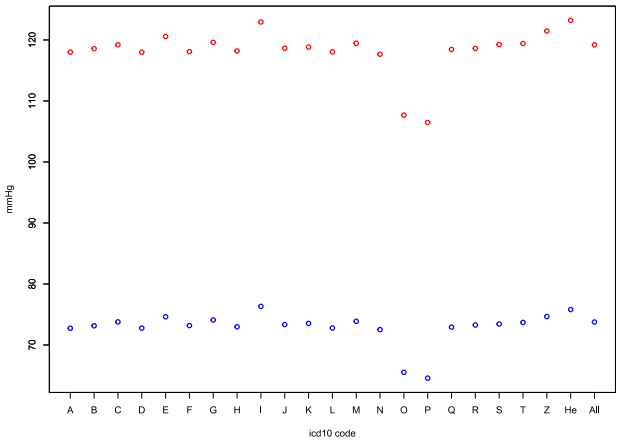

図 4 傷病グループごとの拡張期/収縮期血圧の平均値
Fig. 4 The mean values of diastolic blood pressure and systolic arterial pressure for disease groups.

で後述する健康集団，“*All*” は健康診断データ全体を示している[*3]．傷病グループ O（妊娠，分娩および産じょく）と傷病グループ P（周産期に発生した病態）の個人は A1（非肥満・リスクなし）の割合が飛びぬけて高く，どちらも 6 割を超えているため，健康な個人が多い．傷病グループ E（内分泌，栄養および代謝疾患）や傷病グループ I（循環器系の疾患）は A4（非肥満・服薬投与），B4（肥満・服薬投与）の割合が他グループより高いため，不健康な個人が多い．また，**図 4** に傷病グループごとの拡張期/収縮期血圧の平均値を示す．この結果からも，平均血圧が低い健康なグループ（O，P）と，平均血圧が高い不健康なグループ（E，I）を観測できる．

2.4 健康診断データのクレンジング

健康診断データには多くの欠損値が含まれており，全体の 23.8%のセルが情報を持たないセルである．そのため，分析の前にデータをクレンジングする必要がある．分析の

[*3] 傷病グループ X に属する個人は健診ランクの情報を持たなかったため，省いている．

表 5 クレンジング後の健康診断データの統計量
Table 5 The statistics of the medical examination data after data cleansing.

	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

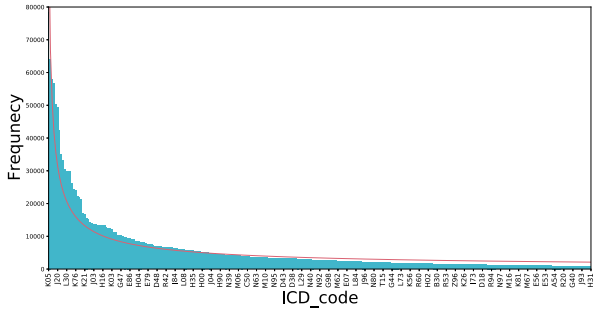

図 5 疾病ごとの罹患者数
Fig. 5 The number of patients for diseases.

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う．

(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 [17] 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．

また，3.4.1 項では，分散 0 の属性を削除している．処理前後の健康診断データの統計量を**表 5** に，データから削除したデータを表 3 の index 列に示す．

レセプトデータに含まれる ICD10 の中分類コード（1,490 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．

$$\text{健診受診年} \leq \text{傷病記録年} \leq \text{健診受診年} + 2$$

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．

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 5** に示す．1,490 種類の傷病のうち 16%（225 種類）はクレンジングによって記録が消え，レコード数が 0 であった．図中の赤線は zipf の法則による近似値である．

3. データ分析

3.1 概要

本章では，次の方法により匿名加工情報の分析を行う．

(1) 健康診断データと，診断されたことのある傷病/処方されたことのある傷病/医薬品のレセプトデータをクロス集計して，疾病の相対リスクを分析する．
(2) がんと脳卒中を対象とし，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

表 6 高血圧の相対リスクに関する 2×2 分割表
Table 6 An example of 2×2 frequency table for the relative risk for high blood pressure.

	A に罹患している	A に罹患していない
高血圧	100	100
正常域血圧	10	190

表 7 高血圧を危険因子とした各傷病の相対リスク $RR_{高血圧}$
Table 7 The relative risk for high blood pressure for diseases.

分類コード	分類	相対リスク
I	循環器系の疾患	1.748
Z	健康状態に影響をおよぼす要因および保健サービスの利用	1.462
E	内分泌，栄養および代謝疾患	1.305
G	神経系の疾患	1.136
C	新生物<腫瘍>血液および造血器の疾患ならびに免疫機構の障害	1.104
T	損傷，中毒およびその他の外因の影響	1.089
M	筋骨格系および結合組織の疾患	1.089
S	損傷，中毒およびその他の外因の影響	1.059
Q	先天奇形，変形および染色体異常	1.059
K	消化器系の疾患	1.000
R	症状，徴候および異常臨床所見・異常検査所見でほかに分類されないもの	0.993
B	感染症および寄生虫症	0.990
D	新生物<腫瘍>血液および造血器の疾患ならびに免疫機構の障害	0.984
J	呼吸器系の疾患	0.973
N	尿路性器系の疾患	0.957
F	精神および行動の障害	0.951
H	眼および付属器の疾患，耳および乳様突起の疾患	0.943
L	皮膚および皮下組織の疾患	0.930
A	感染症および寄生虫症	0.904
O	妊娠，分娩および産じょく	0.184
P	周産期に発生した病態	0.108

て分析して，従来のコホート研究結果と比較する．

(3) 4 種類の予測アルゴリズムを使い，3 年以内の罹患を予測するモデルを 274 種類の傷病について作成する．

3.2 傷病の相対リスク分析 (1)

3.2.1 分析手法

傷病/医薬品グループの相対リスクを求める．相対リスク [12] とは，ある危険因子（たとえば「高血圧」）に曝露した場合，それに曝露しなかった場合に比べて何倍疾病に罹りやすくなるかを表す指標である．例として，**表 6** の場合を考える．高血圧である個人が傷病 A に罹患する確率が 100/200 であるのに対し，高血圧でない個人の罹患率は 10/200 であるため，この場合の相対リスク $RR_{高血圧}$ は $Pr[A|\text{高血圧}]/Pr[A|\text{正常域血圧}] = (100/200)/(10/200) = 10$ である．これは，高血圧の個人はそうでない個人の 10 倍傷病 A にかかりやすい，ということを意味している．

3.2.2 分析結果

高血圧を危険因子とした各傷病の相対リスクを，それぞれ**表 7** に示す．ここで，相対リスクを求める際には罹患年や診断日等の時間情報は無視している．これらの表から，高血圧に対する相対リスクが高いグループ（傷病：I，Z，E）と低いグループ（傷病：O，P）が観測できる．

3.2.3 考察と健康集団

本項では，健康診断データには登場するが，レセプトデータには登場しない個人に着目する．我々はこれらの個人を健康集団 He（傷病を診断されたことも，医薬品を処方されたこともない健康な集団）と呼ぶ．図 2 から分かるように，15,413 人の個人が健康集団に属しており，図 3，図 4 には健康集団 He についての分析結果も示している．

意外なことに，健康集団の診断結果はそれほど健康ではなく，むしろ健診ランクにおいては，図 3 から分かるように A4 や B4 の割合が高く，図 4 から分かるように血圧の平均値も他グループより高い（収縮期：1 位，拡張期：2 位）．健康診断データの他の属性についても分析した結果，健康集団は診断結果はそれほど健康ではないわりに，問診結果は健康的（飲酒はしない，運動はしてる，等）であることが判明した．

3.3 傷病のロジスティック回帰分析 (2)

3.3.1 分析手法

がん（ICD10: C00-C99）と脳卒中（ICD10: I60-I69）を対象にして，3 年以内の罹患を目的変数，健康診断結果を説明変数として，ロジスティック回帰を用いて次のように分析する．

ある被験者 i の 3 年以内の傷病罹患確率 p_{iy} を

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

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

$$z_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M \tag{2}$$

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，$OR = e^{\beta_1}$ で与えられる．罹患数が十分に小さいとき，オッズ比と相対リスクが等しいことがよく知られており [13]，本稿では説明変数 x_1 による罹患影響をオッズ比から確認する．

表 1 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．

3.3.2 分析結果

表 8 に脳卒中，がん，インフルエンザについての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

表 8 ロジスティック回帰結果
Table 8 The result of a logistic regression.

		脳卒中			がん			インフルエンザ	
	特徴量	estimate	OR	RR [1]	estimate	OR	RR [1]	estimate	OR
連続値	const	−3.643 ∗1	0.026		−1.041 ∗2	0.353		−0.494	0.610
	年齢（歳）	0.024 ∗1	1.024	1.14	0.017 ∗1	1.017	1.09	−0.036 ∗1	0.964
	身長（cm）	−0.002	0.998		0.003 ∗4	1.003		0.003 ∗4	1.003
	Body Mass Index（kg/m^2）	−0.004	0.996	1.00	−0.015 ∗1	0.985	0.86	0.009 ∗4	1.009
	収縮期血圧（mmHg）	0.002	1.002	1.02	−0.002 ∗4	0.998	−	−0.001	0.999
	拡張期血圧（mmHg）	0.003 ∗4	1.003		−0.002 ∗4	0.998		−0.005 ∗1	0.995
	中性脂肪（mg/dl）	0.000 ∗2	1.000		0.000 ∗3	1.000		0.000	1.000
	hdl コレステロール（mg/dl）	0.000	1.000	‡	−0.001 ∗3	0.999	0.85	0.000	1.000
	ldl コレステロール（mg/dl）	0.002 ∗1	1.002		−0.002 ∗1	0.998		−0.001 ∗2	0.999
	got ast（IU/L）	−0.001	0.999		0.004 ∗1	1.004		0.003 ∗1	1.003
	γ gtp（IU/L）	0.000	1.000		0.001 ∗1	1.001		0.000	1.000
	hba1c（ngsp）	0.068 ∗3	1.070		0.048 ∗2	1.049		0.072 ∗1	1.074
	赤血球数（$*10^4/\mu$l）	−0.001 ∗1	0.999		−0.001 ∗1	0.999		0.000	1.000
2 値のカテゴリカル	性別	−0.212 ∗1	0.809		−0.530 ∗1	0.588		0.012	1.012
	服薬 1 血圧	0.383 ∗1	1.466	1.56	0.224 ∗1	1.252	1.15	0.127 ∗1	1.136
	服薬 2 血糖	0.249 ∗1	1.283		0.165 ∗1	1.180		−0.116 ∗4	0.890
	服薬 3 脂質	0.255 ∗1	1.291		0.124 ∗1	1.132		0.061 ∗4	1.063
	既往歴 1 脳血管	1.84 ∗1	6.294		−0.117	0.889		0.078	1.081
	既往歴 2 心血管	0.204 ∗2	1.227		−0.025	0.976		−0.079	0.924
	既往歴 3 腎不全・人工透析	0.129	1.138		0.533 ∗1	1.704		0.543 ∗1	1.721
	貧血	0.174 ∗1	1.190		0.227 ∗1	1.255		0.108 ∗1	1.114
	喫煙	0.017	**1.017**	**1.27**	−0.084 ∗1	**0.920**	**1.51**	0.062 ∗2	1.064
	体重変化 20 歳からの	0.052	1.053		0.011	1.011		0.071 ∗2	1.074
	運動習慣 30 分以上	−0.022	0.978		−0.06 ∗2	0.941		−0.016	0.984
	歩行または身体活動	−0.093 ∗2	0.911		−0.109 ∗1	0.897		−0.076 ∗1	0.927
	歩行速度	−0.034	0.966		−0.1 ∗1	0.905		−0.036 ∗4	0.965
	体重変化 1 年間	0.129 ∗1	1.137		0.096 ∗1	1.101		0.109 ∗1	1.116
	食べ方 2 就寝前	0.075 ∗3	1.078		−0.006	0.994		0.068 ∗1	1.070
	食べ方 3 夜食間食	0.014	1.014		0.093 ∗1	1.098		0.049 ∗4	1.050
	食習慣	−0.001	0.999		−0.098 ∗1	0.907		−0.067 ∗2	0.935
	睡眠	−0.239 ∗1	**0.787**		−0.118 ∗1	0.889		−0.178 ∗1	0.837
	保健指導の希望	0.075 ∗3	1.078		0.010	1.010		−0.011	0.989
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	−0.096	0.908		−0.085 ∗4	0.918		−0.094 ∗4	0.911
	非メタボ	−0.075	0.928		−0.108 ∗3	0.898		−0.013	0.987
	食べ方 1（早食い等）								
	普通	0	1.000		0	1.000		0	1.000
	速い	0.095 ∗1	1.100		0.086 ∗1	1.090		0.025	1.026
	遅い	0.023	1.023		0.041	1.042		−0.033	0.968
	飲酒								
	時々	0	1.000		0	1.000		0	1.000
	ほとんど飲まない	0.083 ∗3	1.086		0.044 ∗3	1.045		0.018	1.018
	飲酒毎日	−0.032	0.969		−0.032	0.969		0.015	1.015
	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	0.077	1.080		0.046	1.047		0.048	1.049
	対象外	−0.005	0.995		−0.074 ∗4	0.929		−0.124 ∗3	0.884
	積極的支援	0.072	1.075		0.036	1.037		−0.062	0.940
	尿糖								
	＋	0	1.000		0	1.000		0	1.000
	＋＋	−0.186	0.830		−0.039	0.961		−0.244	0.784
	＋＋＋	−0.304 ∗4	0.738		−0.099	0.906		−0.04	0.961
	−	−0.298 ∗3	0.742		−0.046	0.955		0.010	1.010
	±	−0.345	0.708		−0.058	0.943		0.089	1.093
	尿蛋白								
	＋	0	1.000	−	0	1.000	1.44	0	1.000
	＋＋	0.057	1.059		0.063	1.065		−0.232 ∗4	0.793
	＋＋＋	−0.241	0.786		−0.130	0.878		−0.178	0.837
	−	−0.123	0.884		−0.117 ∗3	0.890		−0.065	0.937
	±	0.031	1.031		0.043	1.043		0.014	1.014
	生活習慣の改善								
	改善予定（1 カ月以内）	0	1.000		0	1.000		0	1.000
	改善するつもりである（おおむね 6 カ月以内）	−0.097 ∗3	0.907		−0.024	0.976		0.020	1.021
	改善するつもりはない	−0.181 ∗1	0.834		−0.125 ∗1	0.882		−0.071 ∗3	0.931
	すでに改善に取り組んでいる（6 カ月以上）	−0.053	0.948		−0.002	0.998		−0.022	0.978
	すでに改善に取り組んでいる（6 カ月未満）	−0.031	0.970		−0.005	0.995		−0.001	0.999

∗1：$P < 0.0001$，∗2：$P < 0.001$，∗3：$P < 0.01$，∗4：$P < 0.05$
−：有意な関連がまったく示されたなかったため表示せず [1].
‡：分析モデルに含めなかったため表示せず.

表 9 当該ヘルスケア企業による匿名加工情報への加工手法
Table 9 The detail of the methods for anonymously processed information.

No	19 条規則	加工方法	該当（健康診断，レセプトにおける）
1	特定の個人を識別できる記述等の全部または一部を削除	規則性のない方法で生成された仮 ID に置換	氏名（健康診断，レセプト）
		削除	住所（レセプト）
2	個人識別符号の全部を削除	削除	被保険者記号（健康診断，レセプト） 被保険者番号（健康診断，レセプト）
3	個人情報と他の情報を連結する符号を削除	規則性のない方法で生成された仮 ID に置換	レセプト ID（レセプト）
4	特異な記述等を削除	トップ・ボトムコーディング	健康診断データ（連続量）（健康診断）
5	他の個人情報との差異等の性質を勘案した措置	一般化	傷病名コードを ICD10 コードに変換（レセプト） 医薬品名コードを ATC コードに変換（レセプト） 診療行為コードをコード表用番号に変換（レセプト）
		日単位 → 月単位に変換	健康診断受診日（健康診断）
		削除	医療機関名称（レセプト）

が 0.000 の値は，estimate がきわめて小さい値と基準値を区別するための表記である）．たとえば，脳卒中で睡眠の OR=0.787 から，睡眠が十分にとれている人（睡眠 $x = 1$）はとれていない人（睡眠 $x = 0$）に比べて脳卒中の 3 年以内罹患リスクが 0.787 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 33 因子，インフルエンザには 25 因が有意であった．

また，表 8 の相対リスク RR は，野田ら [1] の研究結果を表す．ただし，BMI は 19 未満をベースとしたときの 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ときの尿蛋白正常（−，±）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，同等の結果が得られていることが分かる．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 $OR = 0.920$ で不整合となっていた．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（歩行または身体活動，睡眠，食べ方 1，飲酒等）で有意な差が見られた．十分な睡眠をとる生活習慣や 1 日 1 時間以上の歩行または身体活動は，3 つすべての疾病でリスクを下げる効果があった．

3.3.3 考察

本分析では，野田らの先行研究との比較を試みた．野田らの研究では Cox 比例ハザードモデルにより 5 年時生存関数を推定し，相対危険度 RR を求めているのに対し，本研究ではロジスティック回帰により，3 年以内の罹患のオッズ比 OR を求めている．手法が異なる理由の 1 つは，本分析で使用したデータは，主に健康で生存している被験者が中心であり，死亡者のデータが少なかったためである．代替として 3 年以内の罹患を目的変数としたが，罹患は死亡と異なり，1 度罹患した後に再度罹患することもあるので，Cox 比例ハザードモデルで生存関数を算出するのが適切でないため，ロジスティック回帰を使用した．本分析により算出した OR と Cox 比例ハザードモデルによる RR は厳密には一致しないが，ともに重要なリスク比の指標として知られており [29]，対象の疾患による死亡には罹患が前提条件になるため，両者はおおむね同じ傾向を示すと考える．

喫煙に関しては先行研究との整合性が得られなかった．その原因として，喫煙についての変化が考えられる．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，その 15 年後の本分析の 2016 年には 30%（男性）であった [14]．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．このよう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．したがって，匿名加工情報により，従来のコホート研究と同等の分析結果が得られたと結論づける．

3.3.4 匿名加工後の性質について

本匿名加工情報は，当該ヘルスケア企業により，**表 9** に従って法的に適切な加工（法 [27] 第 36 条 1 項，規則第 19 条）が行われており，この加工によって生データの持つ性質が劣化している可能性がある．たとえば Maeda ら [24] は，様々な性質を持つデータセットをセル削除による k–匿名化 [25] で加工することにより，加工で失われる有用性の度合いがデータセットの性質によって異なることを示している．

そこで，表 9 の加工が罹患リスクに及ぼす影響を検討する．まず，削除と仮 ID への置換は罹患リスクに影響を与えない．次に，健康診断受診日を月単位に丸める加工は 3 年以内という条件を変える小さな可能性があるが，無視できる確率である．したがって，影響があるのは表 8 の 12 種の連続量のトップ/ボトムコーディングのみである．そこで，いくつかの仮定をおいて加工前の健康診断データを予測し，当該ヘルスケア企業に確認した後に 3.3.1 項と同様の分析を行った．

表 10 に，匿名加工情報と予測された健診データを用いて算出した OR の絶対誤差の統計量を示す．ほとんどすべての説明変数について，両者の OR は一致しており，たと

表 10 匿名加工情報と予測健診データとの OR 絶対誤差の統計量
Table 10 The statistic of the absolute error of OR between the predicted original data and the anonymously processed information.

病名	平均値	標準偏差	最大値	最小値
脳卒中	$2.5 \cdot 10^{-4}$	$4.2 \cdot 10^{-4}$	$1.9 \cdot 10^{-3}$	$5.6 \cdot 10^{-7}$
がん	$2.1 \cdot 10^{-4}$	$2.7 \cdot 10^{-4}$	$1.2 \cdot 10^{-3}$	$3.5 \cdot 10^{-7}$
インフルエンザ	$2.0 \cdot 10^{-4}$	$2.7 \cdot 10^{-4}$	$1.1 \cdot 10^{-3}$	$4.6 \cdot 10^{-7}$

えば，脳卒中についての絶対誤差は，最大で $1.9 \cdot 10^{-3}$，平均で $2.5 \cdot 10^{-4}$ である．この結果より，匿名加工情報による罹患リスクの変化は無視でき，匿名加工情報を用いても，ヘルスケア情報の分析結果として有用性が認められるレベルの品質の結果が得られると判断する．

3.4 疾病罹患予測モデル (3)

3.4.1 分析手法

本研究では，罹患者が 1,000 人以上の 274 種類の傷病を分析対象とする．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，$y_{A04} = f_{A04}(x_{i1}, x_{i2}, \ldots, x_{i38})$ 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．

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 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．また，本分析では高い精度を出すことが目的ではなく，匿名加工によって機械学習の精度がどれだけ変化するかの分析が目的であるため，各モデルのハイパーパラメータは**表 11** のデフォルト値[*4]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 を用いて実装する．

3.4.2 分析結果

図 6 に 274 種類の疾病の罹患を予測した 4 種類のモデルの F 値の分布を示す．**表 12** に，学習手法ごとの各 274 種類の傷病予測モデルの F 値の統計量を示す．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 6 から傷病の種類により精度の偏りがあることが分かる．**表 13** 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．各学習手法の列は，中分類の平均 F 値を表す．表 13 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ほかに比べて精度が高く，F 値が 0.7 であることが分かる．

表 14 にランダムフォレストの F 値の上位 10 件の疾病を示す．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 [18] であるがん，心疾

[*4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, sklearn.tree.DecisionTreeClassifier.html, sklearn.neighbors.KNeighborsClassifier.html, sklearn.svm.SVC.html
デフォルト値だけでの実験は一般的でないので，今後の実験で最適なパラメータを検討する．

表 11 予測モデルのハイパーパラメータ
Table 11 The hyper parameters of the prediction model.

学習方法	パラメータ名	デフォルト値
k 近傍法	n_neighbors	5
	weights	uniform
	algorithm	auto
	leaf_size	30
	p	2
	metric	minkowski
	C	1
	kernel	rbf
SVC	degree	3
	gamma	scale
	criterion	gini
	splitter	best
決定木	min_samples_split	2
	min_samples_leaf	1
	min_weight_fraction_leaf, min_impurity_decrease	0
	min_impurity_split, ccp_alpha, min_weight_fraction_leaf	
	n_estimaters	100
	criterion	gini
	min_samples_split	2
	min_samples_leaf	1
	max_features	auto
ランダムフォレスト	min_impurity_decrease, verbose, ccp_alpha	0
	bootstrap	TRUE
	oob_score, warm_start	FALSE

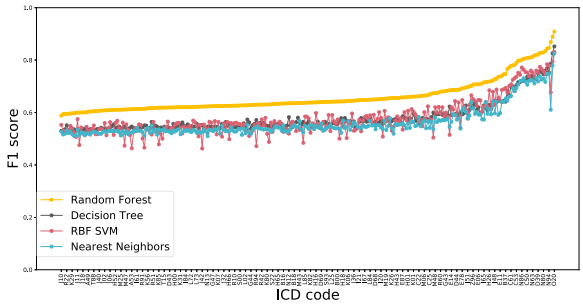

図 6 疾病予測モデルの F 値
Fig. 6 The F value of models predicting diseases.

表 12 各学習手法精度（F 値）の統計量
Table 12 The statistics of accuracy for each of predicting models.

	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 13 ICD10 大分類での平均精度
Table 13 The mean accuracy for ICD 10 blocks.

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	異常検査所見でほか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	健康状態に影響を及ぼす要因等	1	0.707	0.627	0.651	0.616

表 14 F 値上位 10 件の疾病
Table 14 The top 10 diseases in F value.

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	腟および外陰のその他の炎症	11,608	0.827	0.757	0.774	0.738

表 15 日本 3 大死亡原因の罹患予測精度
Table 15 The accuracies of prediction of three major causes of death in Japan.

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

患，脳血管疾患（脳卒中も含む）に該当する傷病の予測精度を**表 15** に示す．脳梗塞は少なくとも 65%の精度で予測可能である．

3.4.3 考察

表 14 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，F 値の上位 10 件中 9 件の傷病であった．原因として，必ず罹患していない患者（男性）を簡単に分類できて，結果的に F 値が高くなったと考えられる[*5]．たとえば，女性データのみで作成した N97 に関するランダムフォレストによるモデルでは，F 値は 0.80 で全データを使用したモデルから約 0.10 劣化した．また，傷病によって精度が異なる原因には，T14（部位不明の損傷）等，健康診断とあまり関係がない傷病があるためと考える．

[*5] N97 のモデルでは，3.4.1 項の手順から性別属性を削除するが，他の属性（年齢，bmi 等）から性別の推定は容易である．

4. *k*–匿名化と分析結果への影響

4.1 概要

健康診断データやレセプトから得られる病歴データを，代表的な匿名化手法である k–匿名性を満たすように加工することにより，3 章で提案した有用性指標がどれだけ変化するのかを明らかにする．k–匿名性は Sweeny によって提案された匿名性の指標 [8] であり，同じ QI を持つ個体の少なくとも k 人が同じ値を持つようにデータを加工すればこれを満たすことができる．k–匿名性を満たすには，レコード等の削除や値の一般化，およびそれの組合せが用いられる．各々の例として，本稿では**表 16** の 2 種類の方法（レコード削除加工，Mondrian アルゴリズム）を評価する．

なお，匿名加工情報に追加の加工を加えても，適法な匿名加工情報と見なせるので，この章では当該ヘルスケア企業による匿名加工情報を「ヘルスケア企業による匿名加工情報」，それらを k–匿名性を満たすように加工したものを「追加匿名加工情報」と区別して呼ぶ．

4.2 QI=性別と年齢

4.2.1 分析手法

本分析では，年齢と性別のそれぞれの値を QI として，ヘルスケア企業による匿名加工情報の健康診断データを表 16 に示す 2 種類のアルゴリズムにより k–匿名化する．それぞれの加工アルゴリズムによって $k = 3, 5, 10, 30, 50, 100, 500, 1{,}000$ で加工したときの追加匿名加工情報に対して，3.4.1 項と同様の分析を行いモデルの精度を比較する．

表 17 に，レコード削除による追加匿名加工情報の k の値による，レコード数と予測精度の変化を示す．各学習手法の値は，274 種類の傷病を予測した際の F 値の平均を表す．$k = 3$ から 1,000 の匿名化を行うと，最大で 10%のレコードが削除され，274 種類の傷病の $k = 0$ の基準データに対する平均 F 値の最大誤差は SVM の 0.007 であった．

また，Mondrian アルゴリズムによる追加匿名加工情報の k の値による予測精度の変化を**表 18** に示す．$k = 3$ から 1,000 の匿名化を行うと，274 種類の平均 F 値の最大誤差は RF の 0.025 であり，レコード削除による追加匿名加工情報よりも誤差が大きくなった．

4.2.2 考察

レコード削除による追加匿名加工情報では，QI を年齢と性別にしたときに，$k = 1{,}000$ でレコードの 10%を削除したが，機械学習の精度は最大でも SVM による 0.007 の劣化であった．この原因の 1 つとして，年齢以外の要素が罹患予測に作用していたことが考えられる．たとえば，ランダムフォレストの高血圧症（I10）の罹患推定における，特徴量重要度 [19] は年齢が 0.06 であるのに対して，収縮期

表 16 k–匿名化手法の詳細
Table 16 The detail of 2 methods for k–anonymity.

	本稿（レコード削除）	Mondrian [30]
方法	該当する人数が k 人未満の QI（年齢，性別）を持つ個人を削除する	QI（年齢，性別）をもとに分割された各グループの QI の値を中央値に置き換える処理を，k–匿名性を満たすまで繰り返す
実装	python による独自実装	Nithin による python スクリプト [31]

表 17 レコード削除によって k–匿名化された追加匿名加工情報の精度
Table 17 The accuracy of k–anonymized medical examination data by the record suppression.

k	レコード数	削除割合	RF	Tree	SVM	KNN
0	203,521	0.0000	0.659	0.579	0.578	0.562
3	203,521	0.0000	0.659	0.579	0.578	0.562
5	203,521	0.0000	0.659	0.579	0.578	0.562
10	203,521	0.0000	0.659	0.579	0.578	0.562
30	203,474	0.0002	0.659	0.579	0.578	0.562
50	203,311	0.0010	0.659	0.579	0.578	0.562
100	202,807	0.0035	0.659	0.579	0.577	0.562
500	196,719	0.0334	0.658	0.578	0.576	0.561
1,000	181,981	0.1058	0.656	0.576	0.571	0.558
最大誤差	–	–	0.003	0.003	**0.007**	0.004
平均誤差	–	–	0.001	0.001	0.001	0.001

表 18 Mondrian アルゴリズムによって k–匿名化された追加匿名加工情報の精度
Table 18 The accuracy of k–anonymized medical examination data by the Mondrian algorithm.

k	RF	Tree	SVM	KNN
0	0.659	0.579	0.578	0.562
3	0.634	0.562	0.567	0.553
5	0.635	0.562	0.567	0.553
10	0.634	0.562	0.567	0.553
30	0.635	0.562	0.567	0.553
50	0.634	0.562	0.567	0.553
100	0.634	0.562	0.567	0.553
500	0.635	0.562	0.567	0.553
1,000	0.634	0.562	0.567	0.553
最大誤差	**0.025**	0.017	0.011	0.009
平均誤差	0.025	0.017	0.011	0.009

血圧が 0.117，BMI が 0.05 と同等もしくはそれ以上であった．したがって，年齢の属性なしでも BMI 等の属性が精度を補償していると考える．

図 7 に健康診断データの年齢頻度と，各年齢における高血圧症の罹患者数を示す．レコード数の少ない 10 代から 20 代や 65 歳以上では，罹患者数がきわめて少ないことが分かる．したがって，k–匿名化をしてそれらのレコードが削除されても精度が最大でも 0.007 の劣化であったことも，理由の 1 つと考える．

また，Mondrian アルゴリズムによる追加匿名加工情報の誤差は，最大，平均ともにレコード削除による追加匿名加工情報よりも大きかった．Mondrian アルゴリズムではデータの QI の値をグループの中央値に書き換える処理を行うため，値を書き換える処理をしないレコード削除加工よりも誤差が大きくなったこと考えられる．k–匿名化には，ほかにも文献 [25] 等の様々な方式が知られているが，上記の理由から，おおむね同様の精度に収まることが予測される．

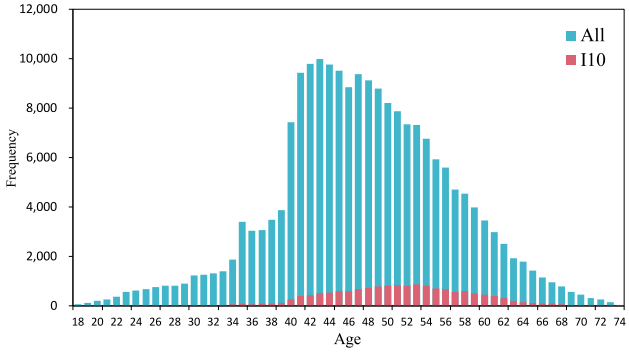

図 7 健康診断患者全体（All）と高血圧症罹患者（I10）の年齢分布
Fig. 7 The distributions of all patients (All) and patients with high blood pressure (I10).

したがって，これらの結果より，レコード削除や一般化によって k–匿名化を行って匿名加工情報を作成したとしても，機械学習の精度に対して重大な影響を与えるほどの低下をしないと本稿では結論づける．

4.3 QI=病歴/処方歴

分析した傷病/医薬品の相対リスクを追加匿名加工情報の有用性と見なし，これらを評価する．

4.3.1 安全性：病歴/処方歴の一意性

傷病/医薬品レセプトデータから得られる病歴/処方歴の一意性に注目し，個人識別リスクを評価する．レセプトデータには 1 顧客についてのレセプトが複数枚分記録されている．それをまとめて各傷病/医薬品について 2 値のベクトル $\boldsymbol{x} = (x_1, \ldots, x_\ell)$，

$$x_i = \begin{cases} 1 & (i \text{ 番目の病歴/処方歴あり}) \\ 0 & (\text{なし}) \end{cases}$$

にし，これを各個人の病歴/処方歴ベクトルとする．

図 8 に，傷病レセプトデータの各個人と同じ病歴を持つ個人数の累積分布を示す．傷病レセプトデータの場合，最大で 5,131 人の個人が同じ病歴（傷病 K：消化器系の疾患のみに罹患したことがある）を持っており，一意な病歴を

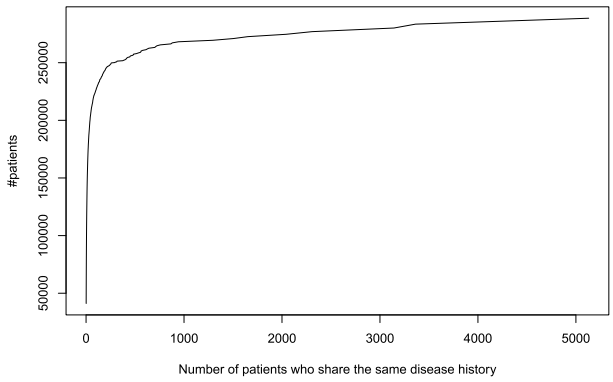

図 8 各個人と同じ病歴を持つ個人数の累積分布

Fig. 8 The distribution of the number of patients who share the same disease history.

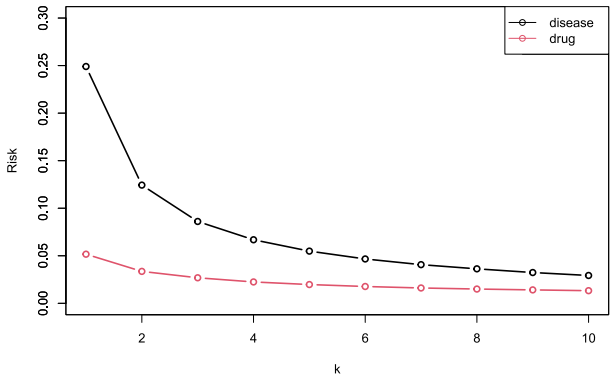

図 9 k–匿名化された病歴（disease）/処方歴（drug）の識別率

Fig. 9 The rate of identified patients in k–anonimized disease/drug history data.

持っている個人は 41,099 人である．$n = 2.8 \cdot 10^5$, $\ell_{病歴} = 23$ のとき，一様ならば各病歴の発生確率は $p = 2^{-23}$ になるため，平均で $n \cdot p = 0.03$ 人の病歴が同じになる．しかし，本データでは平均 283 人の病歴が同じであるため，特定の病歴に著しく偏っていることが分かる．

4.3.2 加工による安全性/有用性の変化

病歴/処方歴は一意な値を持つ個人が多く，これらの人数を減らすために，データ削除による k–匿名化を検討する．

k–匿名化（$k = 1, \ldots, 10$）された病歴/処方歴からの識別率を**図 9** に示す．ここでいう識別率は，元データをすべて持っている最大知識攻撃者 [15] が k–匿名化された病歴から再識別するときの，(識別される人数の期待値)/(加工データに含まれる人数) とする．自分を含めてたかだか k 人と同じ病歴/処方歴を持つ個人の数を n_k，病歴/処方歴に該当する個人数の最大値を n_{max} とすると，$\sum_{k=1}^{n_{max}} n_k/k$ で求めることができる．たとえば $k = 1$ の病歴からは全体の 24.9%（71,864 人/288,568 人）の個人が識別されるが，$k = 10$ になるように該当人数が 10 人未満の病歴を持つ個人 132,736 人を削除すれば，識別される個人割合を 2.9%（4,563 人/155,832 人）まで減らすことができる．最大知識攻撃者は非常に強い仮定であり，その仮定の下での識別率 2.9%は受容可能な範囲である[*6]．

表 19 病歴/処方歴が k–匿名化されたときの傷病/医薬品間の順位相関 ρ

Table 19 The rank correlation ρ between injuries/medicines of k-anonymized data.

k	傷病レセプト	医薬品レセプト
1	1.000	1.000
2	0.996	0.999
3	0.989	0.998
4	0.982	0.998
5	0.976	0.998
6	0.969	0.997
7	0.962	0.997
8	0.958	0.997
9	0.953	0.997
10	0.949	0.996

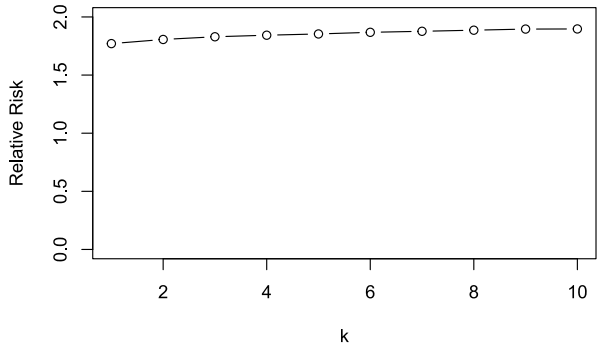

図 10 病歴が k–匿名化されたときの相対リスクの変化

Fig. 10 The distribution of relative risk of high blood pressure of k–anonimized disease history data.

次に，これらの追加匿名加工情報の有用性を評価する．病歴/処方歴は有用なデータであるため，それを評価する指標は数多く想定できるが，ここでは代表として，(1) 傷病/医薬品間の順位相関，(2) 高血圧を危険因子としたときの傷病 I の相対リスクの 2 点で有用性を評価する．

病歴/処方歴を k–匿名化した際に，各分類コード間の順位相関がどの程度変化するのかを，スピアマンの順位相関係数 ρ で評価した結果を**表 19** に示す．レコードを削除して k–匿名化をした場合でも，病歴・処方歴ともに順位相関係 ρ はあまり変化せず，10–匿名化をしても病歴の場合は 0.949，処方歴の場合は 0.996 までしか下がっておらず，データの有用性は失われていない．

また，高血圧を危険因子としたときの傷病 I の相対リスクが，病歴をレコード削除により k–匿名化した際にどのように変化するかを**図 10** に示す．ヘルスケア企業による匿名加工情報の相対リスクが 1.77 であるのに対し，10–匿名化された追加匿名加工情報の相対リスクは 1.90 まで上

[*6] 匿名加工コンテスト PWSCUP2016 [26] では，攻撃者想定として最大知識攻撃者モデルが採用されているが，優勝チームの匿名加工データでさえ 22%の顧客が識別されている．

がっている．これは相対誤差で $(1.90-1.77)/1.77=0.073$ であり，これは野田らの求めた，高血圧治療ありに対する脳卒中の相対リスクの95%信頼区間の幅（0.71）より十分に小さいため，受容可能な精度である．これら2つの有用性評価結果より，追加匿名加工情報が有用であると結論付ける．

5. 先行研究

Arafa ら [20] は，Japan Public Health Center（JPHC）研究プロジェクトの多目的コホートを用いて，89,000 人の成人男女のデータを調査し，尿路結石の既往症の有無による心疾患のリスクを明らかにした．Cox 比例ハザードモデルにより定量化した尿路結石によるハザード比は，1.04 であり統計的な有意水準には達しないことが分かった．また，Islami ら [21] は，2011 年から 2015 年に米国で罹患したがん患者のデータ（the United States Cancer Statistics（USCS）Public Use database）を調べ，全米の州における肥満によるがん罹患のリスクの違いを調査した．BMI が 5 単位増加することによる相対リスク RR は，1.31（胃がん），1.59（肝臓がん），1.10（乳がん）であり，南部や中西部の州では特に高いリスクを検出している．Saint-Maurice ら [22] は，4,840 人の平均 56 歳の成人男性の活動を加速度計で 7 日間測定し，10 年間にわたる追跡コホート調査により，平均歩数が死亡率に及ぼす影響を定量化している．Cox 比例ハザードモデルにより，4,000 歩を基準（$=1$）としたときの 8,000 歩，12,000 歩の被験者の死亡率が単調に下がり，ハザード比でそれぞれ 0.49 と 0.35 になることを示した．Chen ら [23] は，1999 から 10 年間の NHANES（National Health and Nutrition Examination Survey）データを用いて，3 万人の米国成人の栄養補助食品（dietary supplement）の利用有無による，心疾患とがんの死亡率の変化を調査している．適量なサプリメントの取得は死亡率を下げるが，カルシウムを過剰に摂るとがんによる死亡率が調整済みリスク比で，1.62 に増加することを明らかにしている．

このように，コホート研究においては死亡を目的変数とした Cox 比例ハザードモデルによる分析が主流である．罹患を目的変数としてロジスティック回帰分析を用いている研究 [21] は少ない．

6. 個人情報の取扱に対する配慮について

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益な知見を得ることを目的に，匿名加工情報（法 [27] 第 2 条 9 項）を用いている．同法，関連する法令，ガイドライン等を遵守して，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．

厚労省ガイドライン [16] 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が収集した 20 万人分の健康診断データと 28 万人分の傷病/医薬品レセプトデータを分析した．これらのデータはいずれも当該ヘルスケア企業によって適切に匿名加工されたものであるが，匿名加工情報を用いても，ヘルスケア情報の分析結果として有用性が認められるレベルの品質の結果が得られるかどうかを明らかにした．

我々は，相対リスクを用いて傷病/医薬品グループ間の違いを調査することによってデータの有用性を評価した．高血圧を危険因子としたときの傷病 I の相対リスクが 1.77 であることを明らかにした．

飲酒をほとんどしない人が 3 年以内に脳卒中に罹患するリスクは，時々飲酒をする人に比べて 1.09 倍高くなることや，十分な睡眠をとることでリスクを 0.79 倍に下げる等の新たな知見を得た．

健康診断データと傷病レセプトデータから 274 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て評価した結果，ランダムフォレストが最も予測精度が良く，274 種類の傷病の平均 F 値は 0.65 であった．

当該ヘルスケア企業による匿名加工情報の分析結果は，いくつかの仮定をおいて予測された加工前の健康診断データの結果と変わらず，予測健診データとの OR は 38 種の統計量で平均 $2.5\cdot10^{-4}$ の誤差しか生じないことを示した．

さらに，性別・年齢を QI として $k=1{,}000$ までの k–匿名化を行い予測モデルの精度の変化を確認した．$k=1{,}000$ のときレコード数は約 10%減少するが，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．病歴/処方歴を k–匿名化すると，識別される人数の割合は平均 2.9%まで減少するためデータの安全性を高めることができる一方で，相対リスクが相対誤差で 0.073 しか変化しないことを示した．

これらの分析に基づき，匿名加工情報はコホート研究に並ぶ有用なデータであることを確認した．

謝辞 健康診断/レセプトデータをご提供いただいたヘルスケア企業と，倫理的配慮の助言をいただいたコンピュータセキュリティシンポジウム研究倫理相談タスクフォースに感謝する．本研究は JSPS 科研費 JP18H04099 の助成を受けたものです．

参考文献

[1] 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（基本健康検査）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，Vol.53, No.4, pp.265–276 (2006).

[2] 厚生労働省：循環器疾患基礎調査，入手先 ⟨https://www.mhlw.go.jp/toukei/list/junkanki_chousa.html⟩ （参照 2020-11-15）.

[3] NIPPON DATA, 入手先 ⟨https://shiga-publichealth.jp/nippon-data/⟩（参照 2020-08-14）.

[4] 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，Vol.57, No.4, pp.669–673 (2003).

[5] 金子侑紀，小野敦樹，伊藤聡志，菊池浩明，服部充洋，飯田泰興，藤田真浩，山中忠和：匿名加工情報取扱事情者を調査するクローラーシステムの開発，情報処理学会第 82 回全国大会，pp.3_447–3_448 (2020).

[6] 藤田真浩，飯田泰興，服部充洋，山中忠和，松田 規，伊藤聡志，菊池浩明：匿名加工情報取扱事業者による公表情報を利用した匿名加工カタログの提案と実装，コンピューターセキュリティシンポジウム（CSS 2020），pp.1214–1221 (2020).

[7] 株式会社三菱総合研究所：匿名加工情報・個人情報の適正な利活用の在り方に関する動向調査，入手先 ⟨https://www.ppc.go.jp/files/pdf/tokumeikakou_report.pdf⟩（参照2020-11-16）.

[8] Sweeny, L.: *k*-anonymity: A model for protecting privacy, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol.10, No.5, pp.557–570 (2006).

[9] 特定保険組合連合会（けんぽれん）：平成 28 年度 特定検診の「問診回答」に関する調査，入手先 ⟨https://www.kenporen.com/toukei_data/pdf/chosa_h30_08-2.pdf⟩（参照 2020-07-31）.

[10] World Health Organization (WHO): International Statistical Classification of Diseases and Related Health Problems 10th Revision, 入手先 ⟨https://icd.who.int/browse10/2016/en⟩（参照 2020-07-31）.

[11] World Health Organization (WHO): ATC/DDD Index 2020, 入手先 ⟨https://www.whocc.no/atc_ddd_index/⟩（参照 2020-07-31）.

[12] 日本疫学会：疫学用語の基礎知識 相対危険，入手先 ⟨https://jeaweb.jp/glossary/glossary017.html⟩ （参照 2020-07-31）.

[13] 日本疫学会：疫学用語の基礎知識 オッズ，入手先 ⟨https://jeaweb.jp/glossary/glossary019.html⟩ （参照 2020-11-16）.

[14] 国立がん研究センターがん情報サービス：がん登録・統計 喫煙率，入手先 ⟨https://ganjoho.jp/reg_stat/statistics/stat/smoking.html⟩（参照 2020-11-16）.

[15] Domingo-Ferrer, J., Ricci, S. and Soria-Comas, J.: Disclosure risk assessment via record linkage by a maximum-knowledge attacker, *2015 13th Annual Conference on Privacy, Security and Trust* (*PST*), pp.28–35 (2015).

[16] 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）.

[17] 松井秀俊，小泉和之：統計モデルと推測，講談社，p.103 (2019).

[18] 厚生労働省：平成 29 年人口動態統計月報年計の概況，入手先 ⟨https://www.mhlw.go.jp/toukei/saikin/hw/jinkou/geppo/nengai17/index.html⟩（参照 2020-08-14）.

[19] stackoverflow: How are feature_importances in Random Forest Classifier determined?, 入手先 ⟨https://stackoverflow.com/questions/15810339/how-are-feature-importances-in-randomforestclassifier-determined?answertab=votes#tab-top⟩（参照 2020-11-12）.

[20] Arafa, A., Eshak, E.S., Iso, H., Shirai, K., Muraki, I., Sawada, N., Tsugane, S.: for the JPHC Study Group, Urinary Stones and Risk of Coronary Heart Disease and Stroke: the Japan Public Health Center-Based Prospective Study, *Journal of Atherosclerosis and Thrombosis*, Vol.27, No.11, pp.1208–1215 (2020).

[21] Islami, F., Sauer, A.G., Gapstur, S.M. and Jemal, A.: Proportion of Cancer Cases Attributable to Excess Body Weight by US State, 2011–2015, *JAMA Oncol. 2019*, Vol.5, No.3, pp.384–392 (2019).

[22] Saint-Maurice, P.F., Troiano, R.P., Bassett, Jr., D.R., Graubard, B.I., Carlson, S.A., Shiroma, E.J., Fulton, J.E. and Matthews, C.E.: Association of Daily Step Count and Step Intensity With Mortality Among US Adults, *JAMA. 2020*, Vol.323, No.12, pp.1151–1160 (2020).

[23] Chen, F., Du, M., Blumberg, J.B., Ho Chui, K.K., Ruan, M., Rogers, G., Shan, Z., Zeng, L. and Zhang, F.F.: Association Among Dietary Supplement Use, Nutrient Intake, and Mortality Among U.S. Adults: A Cohort Study, *Ann Intern Med.*, Vol.170, No.9, pp.604–613 (2019).

[24] Maeda, W., Shimizu, T., Fukuoka, T. and Morikawa, I.: Dataset Properties and Degradation of Machine Learning Accuracy with an Anonymized Training Dataset, *2020 8th International Symposium on Computing and Networking Workshops* (*CANDARW*), pp.341–347 (2020).

[25] Yamaoka, Y. and Itoh, K.: k-presence-secrecy: Practical privacy model as extension of k-anonymity, *IEICE Trans. Information and Systems*, Vol.100, No.4, pp.730–740 (2017).

[26] 菊池浩明，小栗秀暢，野島 良，濱田浩気，村上隆夫，山岡裕司，山口高康，渡辺知恵美：PWSCUP：履歴データを安全に匿名加工せよ，コンピューターセキュリティシンポジウム（CSS 2016），pp.271–278 (2018).

[27] 個人情報の保護に関する法律（平成 15 年法律第 57 号，平成 27 年法律第 65 号，および，平成 28 年法律第 51 号，令和 2 年法律第 44 号により改正）.

[28] 行政機関の保有する個人情報の保護に関する法律（平成 15 年法律第 58 号）.

[29] 丹後俊郎，古川俊之：医学への統計学第 3 版，朝倉書店，p.195 (2013).

[30] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R.: Mondrian Multidimensional K-Anonymity, *22nd International Conference on Data Engineering* (*ICDE'06*), p.25 (2006).

[31] Nithin Prabhu (Nuclearstar): *k*–anonymity, 入手先 ⟨https://github.com/Nuclearstar/K-Anonymity⟩ （参照 2021-03-26）.

伊藤 聡志 （学生会員）

2017 年明治大学総合数理学部先端メディアサイエンス学科卒業．2019 年同大学大学院先端数理科学研究科先端メディアサイエンス専攻博士前期課程修了．現在，明治大学大学院先端数理科学研究科先端メディアサイエンス専攻博士後期課程在学中．2021 年日本学術振興会特別研究員 DC2.

池上 和輝

2019 年明治大学総合数理学部先端メディアサイエンス学科卒業．2021 年明治大学大学院博士前期課程修了．現在，ソフトバンク株式会社所属．

菊池 浩明 （正会員）

1988 年明治大学工学部電子通信工学科卒業．1990 年明治大学大学院博士前期課程修了．1994 年同博士（工学）．1990 年（株）富士通研究所入社．1994 年東海大学工学部電気工学科助手．1995 年同専任講師．1999 年同助教授，2006 年同情報理工学部情報メディア学科教授．1997 年カーネギーメロン大学計算機科学学部客員研究員．2013 年明治大学総合数理学部先端メディアサイエンス学科専任教授．2016 年明治大学大学院先端数理科学研究科長．2018 年一般社団法人 JPCERT コーディネーションセンター（JPCERT/CC）代表理事．WIDE プロジェクト暗号メールシステム FJPEM の開発，認証実用化実験協議会（ICAT），IPA 独創情報技術育成事業等に従事．暗号プロトコル，ネットワークセキュリティ，ファジィ論理，プライバシ保護データマイニング等に興味を持つ．1990 年日本ファジィ学会奨励賞，1993 年情報処理学会奨励賞，1996 年 SCIS 論文賞，2010 年度，2017 年度情報処理学会 JIP Outstanding Paper Award．2013 年 IEEE AINA Best Paper Award. 2014 年情報セキュリティ文化賞．電子情報通信学会，日本知能情報ファジィ学会，IEEE，ACM 各会員．本会フェロー．