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

伊藤 聡志[1,a)]　池上 和輝[1,b)]　菊池 浩明[2,c)]

概要：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，個人情報の第三者提供とならないデータの匿名加工が必要となった．データを匿名加工する際にはそのデータのプライバシーリスク評価を行う必要がある．本稿では，あるヘルスケア企業が取得した健康診断データとレセプトデータの分析を行い，これらのデータのプライバシーリスク評価を行う．レセプトデータから得られる個人の病歴/処方歴や，健康診断データから得られる個人の身体的特徴や問診表への回答がどの程度一意であるのかを調査し，安全性としてデータから個人が識別されるリスク（個人識別リスク）を評価する．また，2 つのデータを突合することによって，健康診断結果から罹患する病気を予測するために，診断された病気/処方された薬による個人の健康診断結果の違いや傷病/医薬品間の相関関係を明らかにする．

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

Application of Anonymously Processed Information (1): Analysis of Privacy Risk of Aggregation of Medical Examination and Health Insurance Claims

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

Abstract: In Japan, since the Act on the Protection of Personal Information fully came into effect in 2017, a provision of personal processed into the de-identified data to a third party without consent from data principle is established. De-identifying a data requires evaluating a privacy risk of the processed data. In this paper, we analyze and evaluate a privacy risk of a medical examination data aggregated with two kinds of health insurance claims that were collected by a healthcare company. We evaluate an identification risk from the processed data by means of a uniqueness of values and estimate a risk of disease given the medical examination data as an utility of these data.

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

1. はじめに

健康診断は非常に有用なデータであり，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．例えば野田らは，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，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，検査項目と死亡との関係を相対危険度などを用いて明らかにした [1]．しかしながら，2016 年に改正された個人情報保護法では，利用目的を明確にしないで大勢の個人についてのデータをそのまま分析することを禁じており，特

[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)] mmhm@meiji.ac.jp
[b)] cs192021@meiji.ac.jp
[c)] kikn@meiji.ac.jp

に検査結果や病歴などは要配慮情報に分類され，特別な措置を必要とされている．

このプライバシーの問題を解決する手法として，匿名加工情報がある．匿名加工は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であり，匿名加工されたデータから個人を識別しようとする再識別が禁じられている．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，個人情報の第三者提供とならないデータの匿名加工が必要となった．

データを匿名加工するためにはそのデータのプライバシーリスク評価が不可欠であり，リスク評価の研究が国内外で盛んに行われている．匿名加工・再識別コンテスト PWS Cup[2] では，購買履歴データや位置情報データのプライバシーリスクを評価する指標を定め，匿名加工によってそのリスクをどれだけ下げられるかを参加者が競った．また，道廣らは自動車から取得できる移動履歴に注目し，公開されている統計情報や観光スポット情報等を用いて数理モデルを作成し，データの重要な部分のみ抽出する加工を行えば安全性を高めることができることを示している [3]．しかしながら，実際のデータを収集することは困難であるため，こういったプライバシーリスク評価の研究では疑似データや合成データが用いられることが多く，実データのプライバシーリスク評価が求められている．

本稿では，あるヘルスケア企業が取得した 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分析とプライバシーリスク評価を行う．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，10 年間分記録されている．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ある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，実際の匿名加工データからどのような分析結果が得られるかを明らかにするために，本稿ではこれらを無加工のデータとみなして分析を行う．

我々は各データの統計量や頻度分布に加え，以下の 4 点についての分析を行う．

(1) データ中で特異な振る舞いをしている記録またはデータについて報告する．
(2) 診断/処方された傷病/医薬品ごとに個人をグループ分けすると，健康診断結果の平均値や割合がグループごとに大きく異なる場合があり，我々はこのグループごとの差を相対リスクを用いて分析する．
(3) レセプトデータから得られる個人の病歴/処方歴や，健康診断データから得られる個人の身体的特徴や問診表への回答がどの程度一意なものであるのかを調査し，データから個人が識別されるリスクを評価する．
(4) レセプトデータの病歴を加工（k-匿名化 [4]）することによってデータの安全性・有用性がどの程度変化するのかを調査する．

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分析を行い，3 章でこれらのデータの安全性・有用性の評価を行う．

2. データ概要・分析

2.1 健康診断データ

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 1 に示し，各属性に記録されている情報を表 2 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おり，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 [5] への回答結果を示している．

本データには欠損値が多く，全体の 23.8%が {“NaN”, “不明”, “判定不能”, 空欄 } のいずれかの値になっている．表 2 に各属性の欠損値数を示す．最も欠損値が多い属性は第 5 属性の「内臓脂肪面積」であり，情報を持つレコードが 339 しかなく，その一方で第 1,2 属性の「仮個人 id」「診断受診月」に欠損値は存在しなかった．図 1 に，欠損値についてのレコード数分布を示す．横軸が欠損値数，縦軸がレコード数を示しており，例えば 6 つの属性が欠損値であるレコードは約 15 万存在し，49 属性全てが情報を持つレコードは 964,636 レコード中 100 レコードであった．

各レコードには個人 1 名の診断 1 回の結果が記録されており，診断回数は個人によって異なる．図 2 に個人ごとの診断回数のヒストグラムを示す．1 人当たりの平均診断回数は 4.85 回であり，診断回数の最大値は 14 回 (1 人)，最小値は 1 回 (25,812 人)，最頻値は 7 回 (47,763 人) であっ

表 1 3 データの統計情報

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

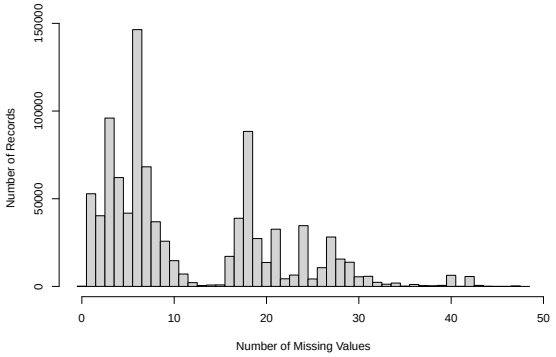

図 1 欠損値を含むレコード数の分布

表 2　健康診断データに記録されている情報

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 slt	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	健康分布	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	既往歴 1 脳血管	350,483	0	$1.20 \cdot 10^{-5}$
32	離散/問診	既往歴 2 心血管	350,393	0	$1.19 \cdot 10^{-5}$
33	離散/問診	既往歴 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	歩行又は身体活動	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	$1.13 \cdot 10^{-5}$

た．一般的に健康診断は 1 年に 1 回だけ受けるものであるが，このデータでは 1 年に 2 回診断している個人が何人か確認できた[*1]．

第 8,9 属性の「収縮期血圧」と「拡張期血圧」について分析する．両方の属性に値を持つのは 810,610 レコードであり，これら 2 属性の値の散布図を図 3 に示す[*2]．x 軸が

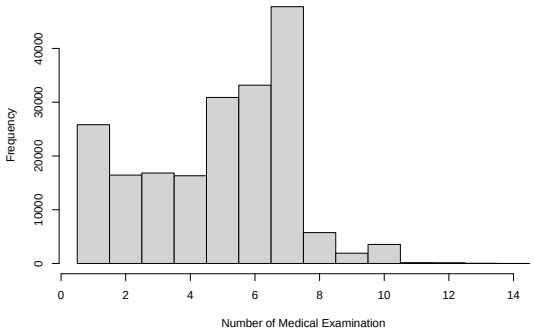

図 2　個人ごとの診断回数のヒストグラム

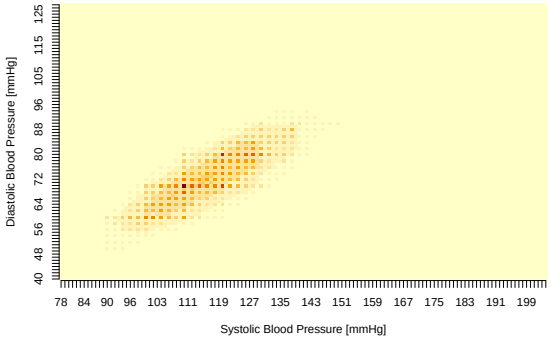

図 3　収縮期血圧と拡張期血圧の分布

収縮期血圧（上の血圧），y 軸が拡張期血圧（下の血圧）を示しており，色が濃いほどその組み合わせが多いことを示している．相関係数は 0.793 であるため，収縮期血圧が高いほど拡張期血圧は高くなりやすい．日本高血圧学会が定めた基準 [6] では，「収縮期血圧/拡張期血圧のどちらか一方，あるいは両方が 140/90mmHg 以上になる病気」のことを高血圧と定義しており，このデータでは全体（レコード）の 13.4%が高血圧に該当する．

第 25 属性の「健康分布」について分析する．健康分布 [8] は，「bmi」や「中性脂肪」等の 12 属性から個人の健康状態にランクをつける指標であり，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康分布の出現（レコード）頻度を表 3 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）に該当し，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．

健康診断データの 49 属性から得られる値の一意性を分析する．表 2 に健康診断データの 2〜49 属性の一意な値の

[*1] 原因として，(1) 個人が健康診断の再検査を受けている，(2) 仮名化した際に 2 名の個人の仮名が重複してしまった，の 2 つが考えられる．

[*2] 図 3 で格子状に色が濃くなっていることからわかるように，収縮期/拡張期血圧は（偶数，偶数）の組み合わせが多くみられるのだが，これはアナログ式の血圧測定器を用いる際は測定値の末尾を偶数にする決まりがある [7] ことが原因だと考えられる．

表 3　健康診断データにおける健康分布の出現頻度

状態	健康分布	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

数を示す．ここでいう「一意な値」とは，その値を持っている個人が 1 名しかいない値のことを指し，例えばデータ中に 5 回登場する値であっても，それらの持ち主が同一人物だったらこれは一意な値である．最も一意な値が多いのは第 5 属性の「内臓脂肪面積」であり，262 種類の値が個人を一意に識別するものであった．「内臓脂肪面積」属性は 7.3〜244.6 が値域である連続値（小数点 1 位）を取り，標準偏差は 43.6 である．

各属性の識別リスクを，属性 V のある 1 つの値 v を背景知識として得た攻撃者により，個人が識別される条件付確率 Pr(一意 |v) の平均値を V による平均識別確率と呼び，この値を表 2 に示す．最も平均識別確率が高いのは第 10 属性の「中性脂肪」であり，評価値は 0.0014 であった．これは，攻撃者がある個人の中性脂肪の情報を得たとき，平均 0.14%の確率で健康診断データからその個人を識別できる，ということを意味している．なお，本稿では欠損値もその属性の値の一つとみなして平均識別確率を求めているため，「内臓脂肪面積」属性の危険度が，一意な値が多い割には低く見積もられている．

2.2　レセプトデータ

本稿で分析する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 1 に示す．

これらのデータで最も重要な情報であると考えられるのが，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 属性にそれぞれ記録されている，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）[9] が，医薬品の分類コードには解剖治療化学分類（ATC 分類）[10] 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．例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，ジギタリス配糖体（心不全の薬）という医薬品は循環器系（大分類コード：*C*）の中の心疾患治療カテゴリ（中分類コード：*C01*）の中の強心配糖体（小分類コード：*C01A*）に含まれるジギタリス配糖体（細分類コード：*C01AA*）に分類される．レセプトデータ中の傷病/医薬品分類コード（大分類）のうち，出現頻度上位 5 件をそれぞれ表 4,5 に示す．

表 4　傷病大分類コードの出現頻度（上位 5 件）

コード	分類	レコード数
K	消化器系の疾患	7,447,520
J	呼吸器系の疾患	6,264,534
H	眼および付属器の疾患，耳および乳様突起の疾患	3,579,158
E	内分泌, 栄養および代謝疾患	3,477,565
L	皮膚および皮下組織の疾患	2,678,177

表 5　医薬品大分類コードの出現頻度（上位 5 件）

コード	分類	レコード数
R	呼吸器系	6,709,468
A	消化管と代謝作用	4,752,345
N	神経系	3,320,980
C	循環器系	3,221,265
J	全身用抗感染薬	2,344,449

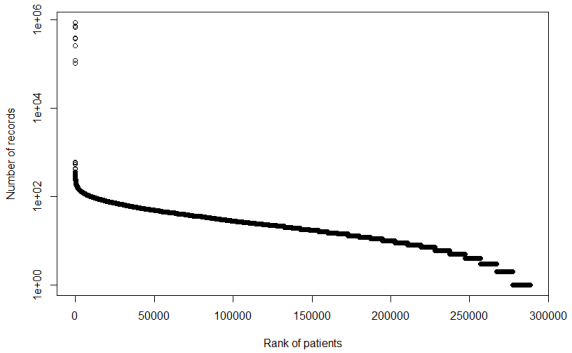

図 4　傷病レセプトデータにおける各顧客ごとのレコード数分布

また，各レセプトデータのレコード数とレセプト数についての分析も行う．表 1 からわかるように，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/枚を 1 レセプトあたりが持っている．しかしこれは一様ではなく，図 4 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．上位 9 名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．

分布図は省略するが，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 名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[*3]．

[*3] 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．

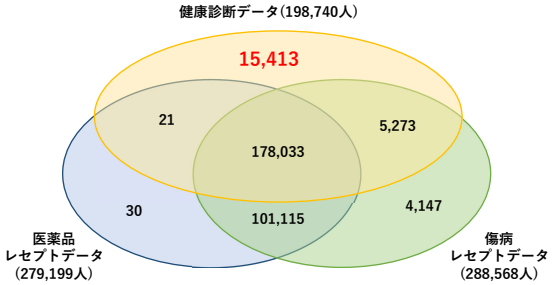

図 5　3 データ間の包含関係

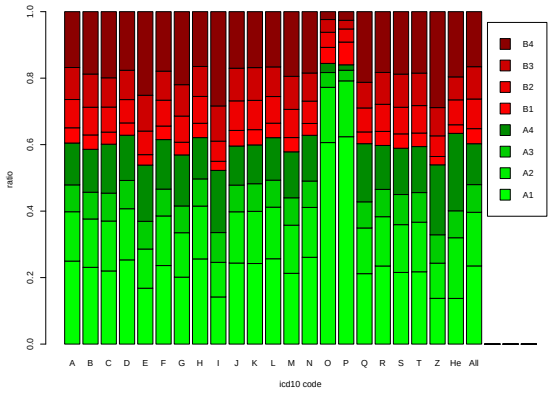

図 6　傷病グループごとの健康分布の割合

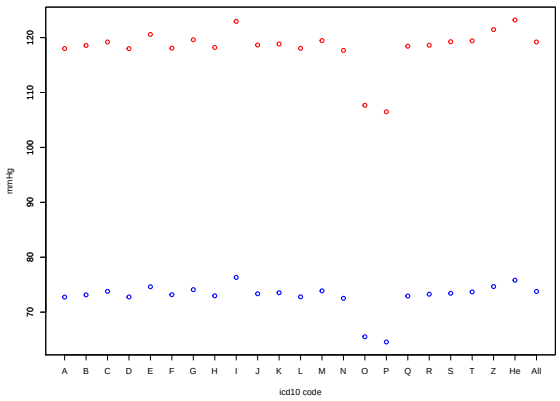

図 7　傷病グループごとの拡張期/収縮期血圧の平均値

2.3　傷病/医薬品グループごとの診断結果の違い

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）は「仮個人 id」属性で結びつけることができる．図 5 に，3 データ間の包含関係をベン図で示す．3 データ全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．

健康診断データに登場する個人を，前述した分類コード（大分類）を用いて，診断されたことのある傷病/処方されたことのある医薬品ごとにクロス集計して分析する．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．例えば今まで傷病 A, B, C だと診断されたことのある個人は，傷病グループ A, B, C 全てに属する．

図 6 に，傷病グループごとの健康分布の割合を示す．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[*4]．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 7 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．

これらの傷病/医薬品グループについて，健康診断結果の相対リスクを求める．相対リスク（relative risk）[11] とは，ある危険因子（本稿では「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 $RR_{\text{高血圧}}$ 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．

表 6　A に関する 2×2 分割表

	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

高血圧を危険因子とした各傷病/医薬品の相対リスクを，それぞれ図 8, 9 に示す．これらの図から，高血圧に対する相対リスクが高いグループ（傷病：I, Z, E，医薬品：C, L, P, T）と低いグループ（傷病：O, P，医薬品：G）が観測できる．このように，健康診断データ・レセプトデータを分析することにより，健康診断結果から罹患リスクの高い傷病/医薬品を予測することができる[*5]．

[*4] 傷病グループ X に属する個人は健康分布の情報を持たなかったため，省いている．

[*5] 傷病/医薬品グループごとに統計量の差があることは，属性推定リスクがあるともいえる．例えば今回の分析の結果から，血圧が高い個人は傷病 I である可能性が高いということがわかってしまう．後述する分類間の相関関係についても同じように扱う．

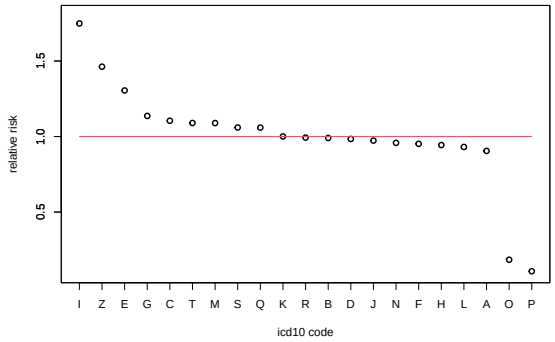

図 8 高血圧を危険因子とした各傷病の相対リスク $RR_{\text{高血圧}}$

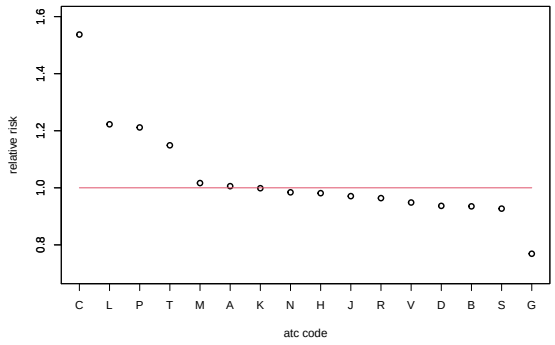

図 9 高血圧を危険因子とした各医薬品の相対リスク $RR_{\text{高血圧}}$

2.4 健康集団

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を，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（健康集団 He）と呼ぶ．図 5 からわかるように，15,413 人の個人が健康集団に属しており，図 6, 7 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，むしろ健康分布においては，図 6 からわかるように A4 や B4 の割合が高く，図 7 からわ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. プライバシーリスク評価

本章では，2 章で分析した傷病/医薬品レセプトデータのプライバシーリスク評価を行う．我々は，病歴/処方歴の一意性をデータの安全性，傷病/医薬品間の相関関係をデータの有用性とみなし，これらを評価する．また，レセプトデータを加工することによって，これらの安全性/有用性がどの程度変化するのかを調査する．

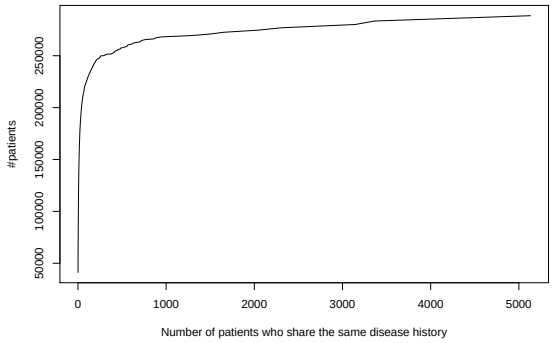

図 10 各個人と同じ病歴を持つ個人数の累積分布

3.1 安全性：病歴/処方歴の一意性

我々は傷病/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．本稿では傷病/医薬品の頻度は考慮しない．例えば傷病が全 3 種類（A, B, C）である場合を考える．ある個人 X はそのうち，傷病 A に 3 回，傷病 C に 2 回罹患したことがあり，傷病 B に罹患したことはないとしたとき，個人 X の病歴ベクトル $\boldsymbol{x}$ は $(1, 0, 1)$ である．傷病/医薬品分類コードの大分類で分析をすると，各個人の病歴ベクトルは $\ell_{\text{病歴}}$ =23 次元，処方歴ベクトルは $\ell_{\text{処方歴}}$ =17 次元になる[*6]．

図 10,11 に，傷病/医薬品レセプトデータの各個人と同じ病歴/処方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，一意な病歴を持っている個人は 41,099 人である．$n = 2.8 \cdot 10^5$, $\ell_{\text{病歴}} = 23$ 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わかる．また医薬品レセプトデータの場合，最大で 2,856 人の個人が同じ処方歴（医薬品 G, L, P, T 以外は処方されたことがある）を持っており，一意な処方歴を持っている個人は 5,226 人であり，平均 378 人の個人が同じ処方歴であった．

3.2 有用性：傷病/医薬品間の相関関係

病歴/処方歴ベクトルからは各分類コード間の相関関係

[*6] レセプトデータには傷病/医薬品コードが空欄であるレコードもあり，今回は空欄も 1 種としてカウントしている．

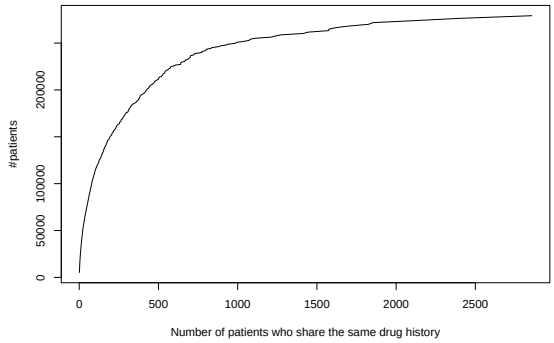

図 11　各個人と同じ処方歴を持つ個人数の累積分布

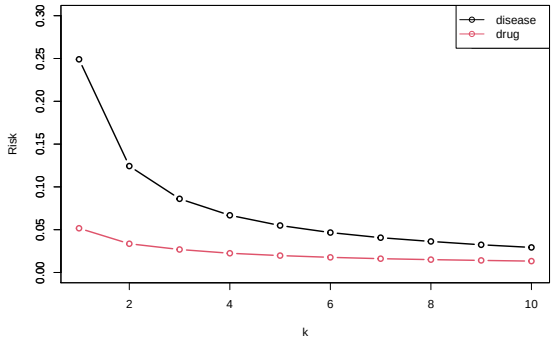

図 12　k-匿名化された病歴 (disease)/処方歴 (drug) の識別率

表 7　傷病における相関関係の高い組み合わせ上位 5 種

組み合わせ	相関係数
傷病 E – 傷病 I	0.441
傷病 D – 傷病 N	0.436
傷病 C – 傷病 D	0.372
傷病 D – 傷病 E	0.368
傷病 E – 傷病 N	0.366

表 8　医薬品における相関関係の高い組み合わせ上位 5 種

組み合わせ	相関係数
医薬品 J – 医薬品 R	0.421
医薬品 A – 医薬品 M	0.344
医薬品 H – 医薬品 K	0.339
医薬品 A – 医薬品 N	0.335
医薬品 J – 医薬品 N	0.328

を求めることができるため，我々はこれをデータの有用性とみなす．表 7, 8 に，傷病/医薬品それぞれにおける相関関係の高い組み合わせ上位 5 種を示す．最も相関係数が高い組み合わせは，傷病では E（内分泌, 栄養および代謝疾患）と I（循環器系の疾患）であり，医薬品では J（全身用抗感染薬）と R（呼吸器系）であった．

3.3　加工による安全性/有用性の変化

病歴/処方歴データの安全性を高めるための加工を考える．3.1 節で分析したように，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，本稿ではデータ削除による k-匿名化を検討する．k–匿名化は Sweeny によって提案された匿名化手法 [4] であり，データ中の少なくとも k 人が同じ値を持つようにデータを加工する手法である．例えば，一意な病歴を持つ個人を全員削除すれば，少なくとも 2 人の個人が同じ病歴を持つようになるので，病歴データを 2–匿名化することができる．

k-匿名化（$k = 1 \ldots 10$）された病歴/処方歴からの識別率を図 12 に示す．ここでいう「識別率」は，元データを全て持っている最大知識攻撃者 [12] 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．自分を含めて高々 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．例えば $k = 1$（無加工）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割合 2.9%は受容可能な範囲である．

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 9 に示す．個人を削除して k–匿名化をした場合でも，病歴・処方歴共に順位相関係 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 k–匿名化した際にどのように変化するかを図 13 に示す．元データの相対リスクが 1.77 であるのに対し，10–匿名化されたデータの相対リスクは 1.90 まで上がっており，相対誤差で $(1.90 - 1.77)/1.77 = 0.073$ であり，加工による有用性の損失が受容範囲であるといえる．

本稿では k–匿名化された病歴/処方歴についての分析をしたが，他のデータの k–匿名化は行っていない．k–匿名化された健康診断データの有用性の変化についての議論は，[13] にて行う．

4.　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（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

表 9　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

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

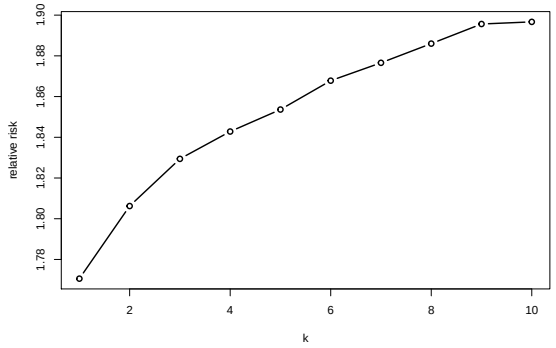

図 13　病歴が k-匿名化されたときの相対リスクの変化

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提供する予定はない．

また，ガイドライン [14] 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．

5. 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ある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，実際の匿名加工データからどのような分析結果が得られるかを明らかにするために，本稿ではこれらを無加工のデータとみなして分析を行った．

我々は，データ中に一意な値（診断結果，病歴/処方歴）がどれくらいあるのかを調査することによってデータの安全性を評価し，また，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．その結果，最大知識攻撃者によって病歴から平均 24.9%の個人が識別されること，分類間の相関係数は平均で 0.142 であること，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．

また，病歴/処方歴を k–匿名化することによる安全性の向上と，有用性の損失を調査した．その結果，病歴を 10–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，分類間の順位相関は 0.051 しか減少せず，また相対リスクも相対誤差で 0.073 しか変化しないことを示した．従って，データの有用性は保たれており，匿名加工された健康診断データ/レセプトデータは安全かつ有用であると結論づける．

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 CSS 研究倫理相談 TF に感謝する．

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，2006 年 53 巻 4 号，p. 265–276.

[2] 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)

[3] 道廣 大喜，長谷川 聡，岡田 莉奈，“自動車の移動履歴データ性質に適した匿名化方式の提案”，暗号と情報セキュリティシンポジウム (SCIS 2020), pp. 1–8. (2020)

[4] L. Sweeny, “k-anonymity: a model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), pp.557–570. (2006)

[5] 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”，`https://www.kenporen.com/toukei_data/pdf/chosa_h30_08-2.pdf`.（2020 年 7 月 31 日参照）

[6] 日本高血圧学会，“一般向け「高血圧治療ガイドライン」解説冊子”，`https://www.jpnsh.jp/data/jsh2014/jsh2014_gen.pdf`.（2020 年 7 月 31 日参照）

[7] 日本循環器病予防学会，“血圧測定手法”，`http://www.jacd.info/method/ketsuatsusokutei.htm`.（2020 年 7 月 31 日参照）

[8] 特定保険組合連合会（けんぽれん），“生活習慣病・健診レベル判定分布とヘルスデータの経年変化に関する調査（平成 26 年 7 月）”，`https://www.kenporen.com/toukei_data/pdf/chosa_h26_7.pdf`.（2020 年 7 月 31 日参照）

[9] World Health Organization (WHO), “International Statistical Classification of Diseases and Related Health Problems 10th Revision”，`https://icd.who.int/browse10/2016/en`.（2020 年 7 月 31 日参照）

[10] World Health Organization (WHO)，“ATC/DDD Index 2020”，`https://www.whocc.no/atc_ddd_index/`.（2020 年 7 月 31 日参照）

[11] 日本疫学会，“薬学用語の基礎知識　相対危険”，`https://jeaweb.jp/glossary/glossary017.html`.（2020 年 7 月 31 日参照）

[12] Josep Domingo-Ferrer, Sara Ricci and Jordi Soria-Comas, “Disclosure Risk Assessment via Record Linkage by a Maximum-Knowledge Attacker”, 2015 Thirteenth Annual Conference on Privacy, Security and Trust (PST), *IEEE*. (2015)

[13] 池上 和輝，伊藤 聡志，菊池 浩明，“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”.（CSS2020 にて発表予定）

[14] 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）.