PWS Cup 2021 – 糖尿病罹患リスクを予測するヘルスケアデータの匿名化コンテスト

菊池 浩明[1,a)]　荒井 ひろみ[2]　井口 誠[3]　小栗 秀暢[4]　黒政 敦史[5]　千田 浩司[6]　中川 裕志[2]
中村 優一[7]　西山 賢志郎[8]　野島 良[9]　波多野 卓磨[10]　濱田 浩気[6]　古川 諒[11]　馬 瑞強[1]
前田 若菜[4]　村上 隆夫[12]　山岡 裕司[4]　山田 明[13]　渡辺 知恵美[14]

概要：健康診断やウェアラブルデバイスから取得したヘルスケアデータは生活習慣病の予測などに活用できる有益なビッグデータである．個人情報取扱事業者は，規則に従った適切な匿名加工に加えて，各種分析の精度を劣化させない最適な加工をすることが求められている．そこで，米国疾病対策予防センター CDC が収集した米国国民健康栄養調査（National Health and Nutrition Examination Survey: NHANES）データを用いて，年齢，学歴，BMI，運動量などの説明変数に対する糖尿病の罹患リスクを正しく評価するための匿名化技術と再識別リスクを探求するコンテストを企画する．

キーワード：プライバシー保護，個人情報，匿名化

PWS Cup 2021 – Competition on Anonymized Healthcare Data to be qualified for Diabetes Prevalence Study

Hiroaki Kikuchi[1,a)]　Hiromi Arai[2]　Makoto Iguchi[3]　Hidenobu Oguri[4]　Atsushi Kuromasa[5]
Koji Chida[6]　Hiroshi Nakagawa[2]　Yuichi Nakamura[7]　Kenshiro Nishiyama[8]　Ryo Nojima[9]
Takuma Hatano[10]　Koki Hamada[6]　Ryo Furukawa[11]　MA Ruiqiang[1]　Wakana Maeda[4]
Takao Murakami[12]　Yamaoka Yuji[4]　Akira Yamada[13]　Chiemi Watanabe[14]

Abstract: Big data from healthcare devices and medical examination are very useful for epidemiologic study predicting a risk of diseases given lifestyle factors. Before sharing de-identified healthcare data, personal data business entities are required to perform an appropriate anonymization algorithm so that it preserves data accuracy and is approved by regulations. In this paper, we design a competition of data anonymization of healthcare data, the National Health and Nutrition Examination Survey, conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention (CDC). The goal of participants is to anonymize the healthcare data to be used to quantify the prevalence of diabetes given demographic characteristics including age, educational level, body mass index, and physical activity.

Keywords: Privacy enhancement, Personal data, anonymization

[1] 明治大学総合数理学部
Meiji University
[2] 理化学研究所
RIKEN
[3] Kii 株式会社
Kii Corporation
[4] 富士通株式会社
FUJITSU LIMITED
[5] 富士通クラウドテクノロジーズ株式会社
FUJITSU CLOUD TECHNOLOGIES LIMITED
[6] NTT 社会情報研究所
NTT Social Informatics Laboratories
[7] 早稲田大学
Waseda University
[8] 株式会社ビズリーチ
BizReach, Inc.
[9] NICT
NICT
[10] 日鉄ソリューションズ株式会社

1. はじめに

本文書の主旨は，コンテスト参加者にその趣旨とルールを正確に伝え，競技を円滑に進めること，および，コンテストで明らかにされる技術や知見の分析を容易にすることにある．米国疾病対策予防センター CDC が収集した米国国民健康栄養調査（National Health and Nutrition Examination Survey: NHANES）データを用いて，年齢，学歴，BMI，運動量などの説明変数に対する糖尿病の罹患リスクを正しく評価するための匿名化技術と再識別リスクを探求するコンテストを企画する．

2. 糖尿病の罹患リスク

2.1 NHANES

NHANES (National Health and Nutrition Examination Survey)[1] は，米国疾病予防管理センター（CDC）の国立衛生統計センター（NCHS）による国民健康栄養調査プログラムとして，1960 年代から行われている調査である．全米 15 箇所で，年 5,000 人をサンプリングの上で調査している．疫学研究，および，健全な公共健康政策やサービスの施策に活用されている．被験者世帯は，NCHS 所長からのプログラムの主旨とプログラム参加依頼のレターを受け取り，同意のもとで個人情報を提供する．被験者には，参加の動機づけとして参加報酬と医師による健康診断結果が後日渡される．

NHANES 2015-2016 の調査プロトコルとデータ収集方法は, NHANES Institutional Review Board (IRB) と NCHS Research Ethics Review Board（ERB）によって承認（プロトコル# 2011 -17[*1]）されている．趣旨に沿ったデータの分析には, 各研究組織での IRB の承認や被験者のインフォームドコンセントは不要である．

2.2 平均活動量と糖尿病罹患率

Zhao らは，NHANES 2015-2016 を用いて，日常的な身体活動量と糖尿病の罹患の関係を明らかにしている [2]．3,932 名の米国の被験者に対して，性別，年齢，人種，教育歴，既婚歴，BMI, 鬱病の有無，貧困状態，身体活動量を説明変数，糖尿病の罹患を目的変数とした多重ロジスティック回帰分析を実施し，身体活動量が最低の被験者セグメント（Q1 = 第 1 四分位数) に対して，活動量最高のセグメント (Q4) は糖尿病の罹患率が 42%に減少することを示した．表 1 に, 各オッズ比， 95%の信頼区間, p 値を示す．

表 1 NHANES 2015-2016 における糖尿病と身体活動量の関係を表すオッズ比 [2] (Model III)

Physical activity	OR	95% CI	p value
Q1	1		
Q2	0.71	0.56 – 0.89	0.003
Q3	0.66	0.52 – 0.84	0.001
Q4	0.58	0.44 – 0.75	< 0.001

目的変数 (outcome) としての糖尿病は，次のように定義している．

(1) 治療薬 (insulin など) を処方されている[*2]．
(2) グリコヘモグロビン $HbA1c \geq 6.5$
(3) 家族に糖尿病患者いるものを除く[*3]
(4) 医師に陽性と診断済み[*4]

説明変数 (exposure) としては，METs (metabolic equivalent score) を用いている [3]．METs は，アンケートの回答をベースとして，活動 [日/週] × 活動時間 [分/日] で定めている．糖尿病患者と健常者の平均 METs は，2,291 min (糖尿病), 3,734 min （非罹患）である．ただし，年齢，教育歴，婚姻歴，BMI, 鬱病の有無，貧困レベルが交絡因子であるとみなし，これらを配慮した複数のモデルで評価をしている．

3. PWS Cup 2021

3.1 ストーリー

健康保険組合 (加工者）は，被保険者の検診データと治療データ（診療履歴）を安全に管理している．被保険者の健康を促進し，治療費を削減するために，糖尿病になるリスクを明らかにしたい．

しかし，被保険者の中には，自分たちのデータが流出していないか懸念のあまり，解析されている匿名化データの中に自分のデータがあるかどうか（メンバーシップ安全性），あるならば，どのデータか（レコード識別性）を推測しようとしているもの（攻撃者）がいる．

しかも，健康保険組合には，適切な匿名化の専門知識がない．そこで，貴方にデータサイエンスの知識を有する委託先として，適切に匿名加工して保険組合を支援して欲しい．

3.2 概要

PWS Cup 2021 は，匿名化フェーズと攻撃フェーズからなる．図 1 に全体的な流れを示す．

加工者 i は，与えられたデータ $B^{(i)}$ に対して，特異なレ

NS Solutions Corporation
[11] NEC
NEC
[12] 国立研究開発法人産業技術総合研究所
AIST
[13] KDDI 総合研究所
KDDI Research, Inc.
[14] 筑波技術大学
Tsukuba University of Technology
[a)] kikn@meiji.ac.jp

[*1] NCHS Research Ethics Review Board (ERB) Approval https://www.cdc.gov/nchs/nhanes/irba98.htm

[*2] NHANES 変数 DIQ050, DIQ070
[*3] DIQ175A
[*4] DIQ160, DIQ010

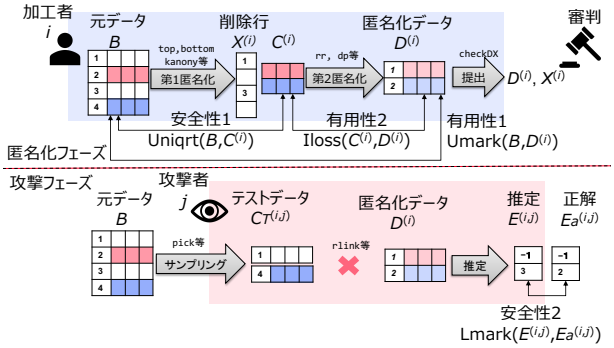

図 1　PWS Cup 2021 処理の流れ

コードを排除する第 1 匿名化と，抽出されたデータを加工する第 2 匿名化を行い，加工データ $D^{(i)}$ と排除レコードの行番号を表す $X^{(i)}$ を提出する．

審判（事務局）は，各チームの $B^{(i)}$ から，攻撃者 j について，排除レコードと抽出レコードを均等にサンプリングして，テストデータ $C_T^{(i,j)}$ を用意する．

攻撃者 j は，$C_T^{(i,j)}$ と，匿名化データ $D^{(i)}$ を与えられて，排除された行番号と，抽出レコードの行番号を推測した $E^{(i,j)}$ を提出する．推測行番号は正解を表す行番号 $E_a^{(i,j)}$ と比較して，3.4 節にて定義される安全性を評価する．

3.3　Data

NHANES 2015-2016 から，表 2, 表 4 の基本統計量を取る 12 属性の 4,190 行のデータを，4,190 行，12 列の行列 $B = (b_{i,j})$[*5]で表す．B の行数を $n = |B| = 4,190$ で表す．列 0, 2, 3, 4, 6, 7, 10 は，それぞれ，性別 (2 値)，人種 (5 値)，学歴 (5 値)，既婚歴 (6 値)，鬱病 (2 値)，貧困 (2 値)，METs 四分位 (4 値）を取る名義変数である．列 1, 5 は，それぞれ，年齢，BMI は連続値を取る．列 11 が目的変数である糖尿病 (1 = 罹患，0 = 非罹患）を表す．

表 3 に，レコードの例を示す．図 2 に，年齢と BMI の分布図を表す．ここで，90%のデータは $BMI < 36$ に分布しているが，それ以上の値を持つ特異なレコードが少数混在していることが分かる．

3.4　競技手順

(1) 加工フェーズ（第 1 匿名化）

加工者 i は，$B^{(i)}$ における特異なデータ（行）を検出し，削除する行を削除行番号リスト $X^{(i)} = (x_1, \ldots, x_{|X|})$ に出力する．$X^{(i)}$ は審判（事務局）に提出する．$B^{(i)}$ から $X^{(i)}$ の行を削除した $n - |X^{(i)}|$ 行，12 列の行列を

$$C^{(i)} = \begin{pmatrix} b_{x_1,0} & \cdots & b_{x_1,11} \\ \vdots & \ddots & \vdots \\ b_{x_{n-|X|},0} & \cdots & b_{x_{n-|X|},11} \end{pmatrix}$$

[*5] 簡単化のために，区別が不要な場合に加工者と攻撃者のインデックスを省略する．

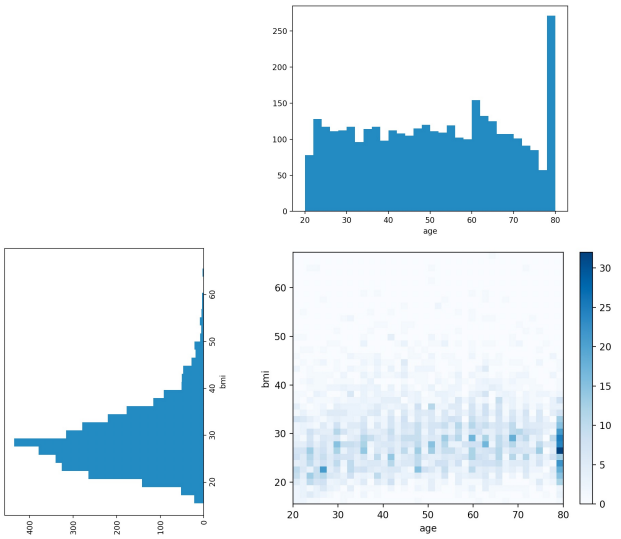

図 2　年齢と BMI の分布

とする．ここで，$x_1, \ldots, x_{n-|X|}$ は，$X^{(i)}$ 以外の $B^{(i)}$ の行番号である．$C^{(i)}$ と $X^{(i)}$ は次の条件を満たさなくてはならない．

(1) 行削除のみを行なう．（値の加工はしない）
(2) しきい値を超えて削除しない．$\theta_n \leq |C^{(i)}|$ を満たす．
(3) $C^{(i)}$ の一意な行数の割合を θ_u 未満にする．$uniqrt(C^{(i)}) \leq \theta_u$.

ここで，一意率 $uniqrt()$ は 3.6.1 節で後述する．

(2) 加工フェーズ（第 2 匿名化）

加工者 i は，再識別が困難になるようにデータ $C^{(i)}$ の値を加工し，匿名化データ $D^{(i)}$ を提出する．

(1) 値の加工のみを行なう．（行削除はしない）
(2) 離散値は，$B^{(i)}$ の値域の値のみを取る．連続値は，$B^{(i)}$ の対応する行の最小値と最大値の区間内の値を取る．
(3) 有用性指標 $Umark(B^{(i)}, D^{(i)})$, $Uniqrt(C^{(i)}, D^{(i)})$ が表 5 の条件を満たす．

有用性指標の定義は 3.5 節にて後述する．ただし，表 5 のしきい値は競技の途中で変更することがあることに注意せよ．

(3) サンプリング

審判（事務局）は，チーム i の $B^{(i)}$ からランダムサンプリングを行い，各攻撃者 j について，テストデータ $C_T^{(i,j)}$ を算出する．

図 3 に，サンプリングの例を図示する．元データ $B^{(i)}$, 削除行番号 $X^{(i)}$, 残存データ $C^{(i)}$ に対して，$X^{(i)}$ から 50 行（負例），$C^{(i)}$ から 50 行（正例）をランダムサンプリングして，$m = |C_T| = 100$ のテストデータとする．匿名化データ $D^{(i)}$ は，$C^{(i)}$ と同じ行数，すなわち，$|C^{(i)}| = |D^{(i)}|$, かつ，$|X^{(i)}| + |C^{(i)}| = |B^{(i)}|$ とする．

また，事務局は，このサンプリングを i の正解行番号 $E_a{}^{(i)}$ に保存し，攻撃者に対して秘匿する．ただし，この時の行番号は，匿名化の際にサンプリングされた範囲で付番

表 2　データセット Diabetes の基本統計量 (離散値)

	0	2	3	4	6	7	10	11
count	4190	4190	4190	4190	4190	4190	4190	4190
unique	2	5	5	6	2	2	4	2
top	Female	White	College	Married	0	0	Q1	0
freq	2117	1398	1214	2171	3313	3306	1168	3317

表 3　データセット Diabetes（一部）

	0	1	2	3	4	5	6	7	8	9	10	11
0	Male	62.0	White	Graduate	Married	27.8	0	0	0	0	Q2	1
1	Male	53.0	White	HighSchool	Divorced	30.8	0	1	0	0	Q1	0
2	Male	58.0	White	HighSchool	Divorced	38.8	0	0	0	0	Q1	1
3	Female	56.0	White	Graduate	Parther	42.4	1	0	0	0	Q3	0
4	Female	42.0	Black	College	Divorced	20.3	1	0	0	0	Q4	0

表 4　データセット Diabetes の基本統計量 (連続量)

	1	5
count	4190.000000	4190.000000
mean	50.455847	29.184010
std	17.887312	6.850947
min	20.000000	14.500000
25%	35.000000	24.400000
50%	50.000000	28.100000
75%	65.000000	32.700000
max	80.000000	67.300000

表 5　しきい値 (仮)

項目	指標	条件
行数	$\|C\|$	$\|C\| \geq \theta_n = n/2$
一意率	$Uniqrt(B, C)$	$uniqrt \leq \theta_u = 0.5$
罹患率	$Umark(B, D)$	$rate \leq \theta_r = 0.05$
相関	$Umark(B, D)$	$cor \leq \theta_c = 0.1$
リスク比	$Umark(B, D)$	$OR \leq \theta_o = 0.1$
情報損失	$Iloss(C, D)$	$Iloss \leq \theta_l = 6$

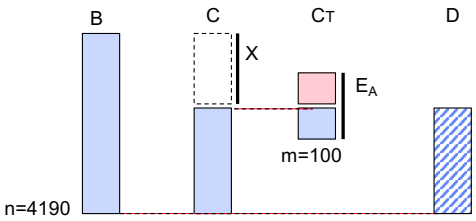

図 3　テストデータ C_T の生成

されることに注意が必要である．表 6 にサンプリングの例を示す．$X = (2, 5, 6, 8, \ldots)$ で排除されている時，サンプリングは表の排除（○）と維持（●）のそれぞれから選ぶ．その時の行番号は,$B^{(i)}$ の番号ではなく，加工データ $D^{(i)}$ の相対行番号とする．削除された行は，値 -1 で表す．この例の時，正解となる行番号は E_a の様になる．

(4) 攻撃フェーズ

攻撃者 j は，加工者 i の匿名化データ $D^{(i)}$ とテストデータ $C_T^{(i,j)}$ が与えられ，$C_T^{(i,j)}$ の全ての行について，削除された特異な行がどれであるか推測し，削除されていなければ $D^{(i)}$ のどの行であるかを再識別する．推定結果は

$$E^{(i,j)} = \begin{pmatrix} e_{x_1,1} & \cdots & e_{x_1,k} \\ \vdots & \ddots & \vdots \\ e_{x_{|X|},1} & \cdots & e_{x_{|X|},k} \end{pmatrix}$$

の $n - |X|$ 行，k 列の行列であり，上位 k 候補を k 個の列に指定する．削除を検出した行は，k 列とも -1 を指定する．

表 7 に推定された行番号 E の例と対応する正解行番号 E_a を示す．正解している推定をゴシックで示す．この例では，2 行目が削除行番号と推測されている．

(5) 判定

審判（事務局）は，加工者 i の攻撃者 j による推定 $E^{(i,j)}$ の正しさを，加工者 i についての正解行番号 $E_a^{(i,j)}$ について次の様に定められた安全性 $recall$, $prec$, $topk$ で評価する．

表 6　テストデータにおけるインデックスの例

B	X	C_T	E_a	D
0				0
1				1
2	2	○	-1	
3		●	2	2
4		●	3	3
5	5			
6	6			
7				4
8	8	○	-1	
9				5

表 7　推定行番号 E の例

推定 $E_{\ell,1}$	$E_{\ell,2}$	$E_{\ell,3}$	正解 E_a	判定
29	847	2599	29	√
-1	-1	-1	-1	√
2038	**2345**	2336	2345	√
2702	1378	2331	80	NG
134	1820	2580	-1	NG

$$recall = \frac{|{E_a}^X \cap E^X|}{|{E_a}^X|}, prec = \frac{|{E_a}^X \cap E^X|}{|E^X|},$$
$$top_k = \frac{|\{\ell \in {E_a}^X | e_{a\ell} \in E_\ell\}}{|E_a^X|}$$

ここで，E^X は，E において > -1（非削除）の（残存する）行の番号の集合，すなわち，

$$E^X = \{\ell \in \{0, \ldots, n-1\} | e_{\ell,1} > -1\}$$

を表す．表 7 の例では，$E_a^X = \{1,3,4\}$, $E^X = \{1,3,4,5\}$ である．従って，この例では，$recall = |\{1,3,4\}|/|\{1,3,4\}| = 3/3$, $prec = 3/4$ である．また，E_ℓ は E の ℓ 行目の全ての列の値の集合 $E_\ell = \{e_{\ell,1}, \ldots, e_{\ell,k}\}$ であり，例 7 の例では，$E_1 = \{29, 847, 2599\}$ である．ここには，正解行番号の $e_{a1} = 29$ を含むので，分子の集合の要素となり，全体では $top_k = \frac{|\{1,3\}|}{|\{1,3,4\}|} = 2/3$. 攻撃の総合危険度を，$recal \times prec \times top_k$ で定める．

(1) 匿名加工部門: 加工者 i に対する推定の総合危険度の全攻撃者の最大値（の低さ）で評価する．

(2) 攻撃部門: 匿名加工部門の上位 3 位の平均危険度（の高さ）で評価する．

(3) 予備戦 1 対本戦 9 の比で評価する．

3.5 有用性指標

3.5.1 罹患率クロス集計 ccount

説明変数のそれぞれの値について，糖尿病と非罹患しているレコード（被験者）の数を集計する．カウント (cnt) と全体における比 (rate) を算出する．ただし，名義変数については，すべての値について，連続量については，いくつかの領域に区分化して集計する．

元のデータと匿名化されたデータの各属性ごとの糖尿病者の比率の最大絶対誤差で有用性を評価する．匿名化データ $D = (d_{\ell,c})$ の列 c が値 v の時，糖尿病である ($diabete = 1$) 度数を $\chi^{(d)}_{c,v}(D)$

$$= |\{\ell \in D^X | d_{\ell,c} = v, d_{\ell,diabetes} = d, \ell \in \{0, \ldots, 7, 10\}\}|$$

で表す．ここで，$diabetes = 11$ 列目である．

表 8 に集計した結果の例（一部）を示す．例えば，B における女性の糖尿病患者数 (cnt) は 407 名，非罹患は 1,710 名であり，[2] の Table 1 とほぼ一致している．その比率 (rate) は $\chi^{(0)}_{0,Female}(B)/n = 0.408$, $\chi^{(1)}_{0,Female}(B)/n = 0.097$ である．

3.5.2 共分散行列 cor

共分散指標は，各列の値の相関を表す．$r_{i,j}(D)$ は，匿名化データ D の共分散行列 $R = (r_{i,j})$ の係数であり，異なる列 i と j 間のピアソンの相関係数を与える．離散値の場合は値 v についてダミー変数に展開する．

表 9 に，共分散行列の値の一部を示す．ここで，下三角行列の値のみを残して対称な要素を 0 にしている．第 0 列（性別）の Male と Female の様に排他的な値は -1, 第 2 列（人種）の様に複数の値がある時は，相関の強さを表す．第 1 列（年齢）は，連続値なので，列全体で他の列との相関を求める．正の相関を赤，負の相関を青のグラデーションで表しており，例えば，`4_Widowed` は，1 列の年齢との間に大きな正 (赤) の相関 0.34 があることが分かる（離婚率は年齢と正の相関がある）．

表 8　クロス集計指標 ccount

	cnt		rate	
diabetes	0	1	0	1
Female	1710	407	0.408	0.097
Male	1607	466	0.384	0.111
(19, 44]	1574	110	0.376	0.026
(44, 64]	1033	379	0.247	0.090
(64, 80]	710	384	0.169	0.092
Black	647	208	0.154	0.050
Hispanic	443	127	0.106	0.030
Mexican	518	200	0.124	0.048
Other	537	112	0.128	0.027
White	1172	226	0.280	0.054
11th	373	120	0.089	0.029
9th	386	168	0.092	0.040
College	975	239	0.233	0.057
Graduate	851	156	0.203	0.037
HighSchool	732	190	0.175	0.045
Divorced	342	103	0.082	0.025
Married	1669	502	0.398	0.120
Never	652	87	0.156	0.021

表 9　共分散行列指標 cor

	0_Male	0_Female	1	2_White	2_Black	2_Mexican	2_Other	2_Hispanic	3_
0_Male	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
0_Female	-1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
1	0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	
2_White	0.02	-0.02	0.12	0.00	0.00	0.00	0.00	0.00	
2_Black	-0.02	0.02	-0.05	-0.37	0.00	0.00	0.00	0.00	
2_Mexican	-0.02	0.02	-0.03	-0.34	-0.23	0.00	0.00	0.00	
2_Other	0.04	-0.04	-0.11	-0.30	-0.20	-0.18	0.00	0.00	
2_Hispanic	-0.03	0.03	0.02	-0.29	-0.19	-0.18	-0.16	0.00	
3_Graduate	0.01	-0.01	-0.06	0.09	-0.07	-0.15	0.19	-0.07	
3_HighSchool	0.03	-0.03	-0.01	0.02	0.07	0.00	-0.09	-0.01	
3_College	-0.06	0.06	-0.05	0.11	0.07	-0.13	-0.06	-0.03	
3_11th	0.04	-0.04	0.01	-0.10	0.02	0.08	-0.02	0.04	
3_9th	-0.01	0.01	0.16	-0.21	-0.10	0.31	-0.04	0.11	
3_nan	-0.02	0.02	0.02	-0.01	-0.01	-0.01	-0.01	0.04	
4_Married	0.10	-0.10	0.14	0.04	-0.15	0.07	0.08	-0.04	
4_Divorced	-0.06	0.06	0.14	0.03	0.04	-0.06	-0.05	0.02	
4_Parther	0.02	-0.02	-0.21	-0.03	0.00	0.02	-0.03	0.05	
4_Separated	-0.03	0.03	0.02	-0.06	0.04	0.01	-0.05	0.07	
4_Never	0.00	0.00	-0.37	-0.08	0.16	-0.06	0.02	-0.03	
4_Widowed	-0.14	0.14	0.34	0.07	-0.01	-0.02	-0.06	-0.01	
5	-0.08	0.08	0.06	-0.04	0.09	0.11	-0.19	0.04	
6	-0.08	0.08	-0.03	0.03	-0.04	-0.02	-0.02	0.04	
7	-0.03	0.03	0.00	-0.21	0.04	0.16	-0.04	0.11	

3.5.3 糖尿病罹患オッズ比 OR

糖尿病の罹患についての多重ロジスティック回帰を行い，各説明変数の値に応じたオッズ比 (OR) を評価する．$OR_{c,v}(B)$ は，列 c の因子 v についての糖尿病罹患の (交絡因子調整済み) 相対リスクであり，

$$OR_{c,v}(B) = \frac{p_1}{1-p_1}\frac{1-p_0}{p_0} = e^{\beta_v}$$

で定められる．ここで，β_v は多重ロジスティック回帰モデルの説明変数 v に対応する係数，p_0, p_1 は因子 v で条件つけられた糖尿病になる確率であり，$p_0 = Pr[b_{\ell,\text{diabetes}} =$

表 10　糖尿病罹患オッズ比 OR

	Coef	OR	pvalue
Intercept	-7.319	0.001	0.000
gen[T.Male]	0.380	1.463	0.000
race[T.Hispanic]	-0.302	0.740	0.062
race[T.Mexican]	0.084	1.088	0.578
race[T.Other]	0.020	1.020	0.909
race[T.White]	-0.844	0.430	0.000
edu[T.9th]	-0.151	0.860	0.404
edu[T.College]	-0.073	0.930	0.652
edu[T.Graduate]	-0.150	0.861	0.395
edu[T.HighSchool]	-0.192	0.825	0.244
mar[T.Married]	0.278	1.320	0.069
mar[T.Never]	0.326	1.386	0.101
mar[T.Parther]	0.316	1.372	0.168
mar[T.Separated]	0.081	1.084	0.779
mar[T.Widowed]	-0.134	0.875	0.515
qm[T.Q2]	-0.228	0.796	0.065
qm[T.Q3]	-0.328	0.720	0.014
qm[T.Q4]	-0.401	0.670	0.004
age	0.057	1.058	0.000
bmi	0.097	1.102	0.000
dep	0.440	1.552	0.000
pir	0.147	1.158	0.174

$1|b_{\ell,c}=v_0]$, $p_1 = Pr[b_{\ell,\text{diabetes}} = 1|b_{\ell,c}=v]$ である．v_0 は，c 列における基準 (コントロール) となる値域の一つである．

表 10 に，ロジスティック回帰の結果を示す．ここで，Coef，OR, pvalue は，それぞれ，係数 β, オッズ比, p 値を表す．連続値は，その値が 1 増えるごとに罹患するリスクの OR を表す．名義変数については，それぞれ最初の値をコントロール v_0 として，そのオッズ比を与える．例えば，性別は $OR_{\text{gen,male}} = 1.43$ は，$v_0 =$ Female に対する $v =$ Male の OR が 1.43，すなわち，男性は女性に対する糖尿病のリスクが 1.46 倍高く，年齢が 1 歳増えるたびに，糖尿病のリスクが 1.058 倍高くなることを表している．人種は，$v_0 =$ Black に対するそれぞれの人種の OR を表している．罹患リスクが増える (OR が 1 以上) 因子を緑，減少する因子を赤で表す．p-value の列は，0.05 を有意水準とした時の有意な p 値を赤で示す．

活動量 METs に対する罹患リスクは，第一四分位数 Q1 に対する第 2(qm[T.Q2])，第 3(qm[T.Q3]) の OR で表されており，定期的に運動するほど糖尿病に罹患するリスクが 0.79, 0.72 と減る．これらは，表 1([2] の Table 2) とほぼ一致している．

3.5.4　有用性指標 U-Mark

罹患比, 共分散, オッズ比を, 元のデータと匿名化データのそれぞれについて算出し, それらの最大絶対誤差 MAE(Max Absolute Error) をもって，匿名化データの有用性（誤差）の値とする．元データ B, 匿名化データ D に対して，

$$
\begin{aligned}
cnt(B,D) &= \max_{c,v,d} \left|\chi^{(d)}_{c,v}(B) - \chi^{(d)}_{c,v}(D)\right|, \\
rate(B,D) &= \max_{c,v,d} \left|\frac{\chi^{(d)}_{c,v}(B)}{|B|} - \frac{\chi^{(d)}_{c,v}(D)}{|D|}\right|, \\
coe(B,D) &= \max_{i\neq j} \left|r^{(v)}_{i,j}(B) - r^{(v)}_{i,j}(D)\right|, \\
OR(B,D) &= \max_{\ell,v} |OR_{\ell,v}(B) - OR_{\ell,v}(D)|,
\end{aligned}
$$

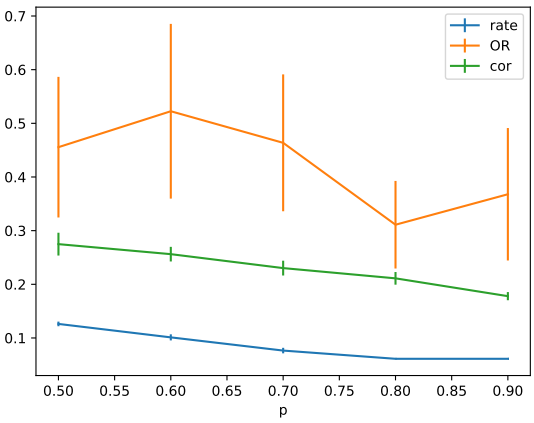

図 4　RR についての有用性指標の変化

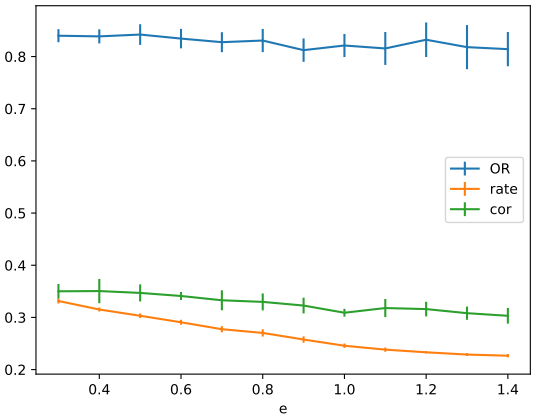

図 5　DP についての有用性指標の変化

と定める罹患率，共分散行率，オッズ比の偏差を U-Mark と呼ぶ．

表 11 に，これらを算出した有用性指標 U-Mark 値の例を示す．図 4 と 5 に，それぞれ，離散値を取る列 (0, 2,3,4,6,7,10) に維持確率 $p = 0.5, \ldots, 0.9$ についてランダマイズドレスポンスした時と，連続値を取る列 (1,5) にプライバシー費用 $\epsilon = 0.3, \ldots, 1.4$ でラプラスノイズを加えた有用性の変化を示す．10 回行った時の標準偏差をエラーバーで示している．

3.5.5　有用性指標 Iloss

加工によって失われた情報量の損失を, Iloss (Information Loss) で定量化する．データ C と D について，

$$
Iloss(C,D) = \max_{u\in\{1,\ldots,|C|\}} d(\boldsymbol{C}_u, \boldsymbol{D}_u)
$$

とする．ここで，$\boldsymbol{C}_u$ を C の u 行のベクトル $\boldsymbol{C}_u = (c_{u,0}\cdots c_{u,11})$, $d()$ は $\boldsymbol{C}_u$ と $\boldsymbol{D}_u$ の L1 距離[*6], $d(\boldsymbol{C}_u, \boldsymbol{D}_u) = \max_v d(c_{u,v}, d_{u,v})$，ただし，

[*6] [5] では，Information Loss の多様な定義を示しているが，本定義とは異なっている

表 11　有用性指標 U-Mark の例

	cnt	rate	Coef	OR	pvalue	cor
max	596.000000	0.065067	0.385540	0.218689	0.455050	0.182816
mean	114.606061	0.010027	0.102278	0.082985	0.129358	0.015871

表 12　Lloss の例

$\boldsymbol{C}$	M	62	Wh	G	Mar	27	0	0	0	0	Q2	1
$\boldsymbol{D}$	M	53	Wh	HS	Div	30	0	1	0	0	Q1	0
d	0	9	0	1	1	3	0	1	0	0	1	1

$$d(c_{u,v}, d_{u,v}) = \begin{cases} |c_{u,v} - d_{u,v}| & \text{if } v \text{ 列が連続値,} \\ |\{v | c_{u,v} \neq d_{u,v}\}| & \text{if } v \text{ 列が離散値,} \end{cases}$$

とする．

表 12 に，(1 行づつの)$\boldsymbol{C}$, $\boldsymbol{D}$ についての，Iloss の計算例を示す．$ILoss(\boldsymbol{C}, \boldsymbol{D}) = \max(|62-53|, |27-30|, |\{3,4,7,10,11\}|) = \max(9,3,5) = 9$ である．

3.6　安全性指標

3.6.1　一意率 Unique Rate

一意率は，一意な行数の割合で定める．極端に大きい値を含むレコードをトップコーディングなどにより削除することにより，(B に対する) 一意率を下げる．元データ B, (第 1) 匿名化データ C における一意率は，

$$uniqrt(B, C) = \frac{|\{C_\ell | \ell = 1, \ldots, |C|\}|}{|B|}$$

と定める．ただし，C_ℓ は ℓ 行目の全列からなる 10 列のタプル, $(c_{\ell,0}, \ldots, c_{\ell,10})$，とし，連続量の列 (1,5) においては，10 の位で丸めた値とする．タプルの集合なので結局一意な行の種類数を与える．分母が C でなく，元データ B であることにも注意せよ．

例えば，表 3 の例の B について，2 行目 (行 1) と 3 行目 (行 2) は同じタプルとなり，$X = (4)$ で，$C = \begin{pmatrix} B_0 \\ \vdots \\ B_3 \end{pmatrix}$ とすると，$uniqrt(B,B) = 4/5$, $uniqrt(B,C) = 3/5$ である．

一意率は，トップコーディング，k-匿名化などにより，一意な行を削除することで高めることが出来る．図 6 に，BMI の値にトップコーディングと離散値の列について k-匿名（行削除）を適用した時の一意率の減少の効果を示す．rows は削除された行数の比率を示している．BMI が高い値は一意であることが多いが，件数が少いために比にはさほど影響しないことが分かる．

3.6.2　所属推定率・再識別率

安全性指標 L-Mark (Link Benchmark) は，所属推定の正しさとレコードリンク攻撃による匿名化された行の再識別の正しさの総合値である．正解行番号 E_a と推定行番号 E について，`lmark.py Ea.csv E.csv` により評価する．

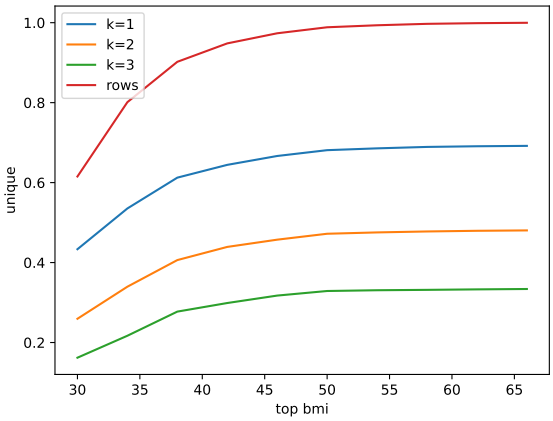

図 6　Top coding と k-匿名による一意率の変化

表 13　安全性評価指標 L-Mark の例

recall	prec	topk
0.84	0.84	0.74

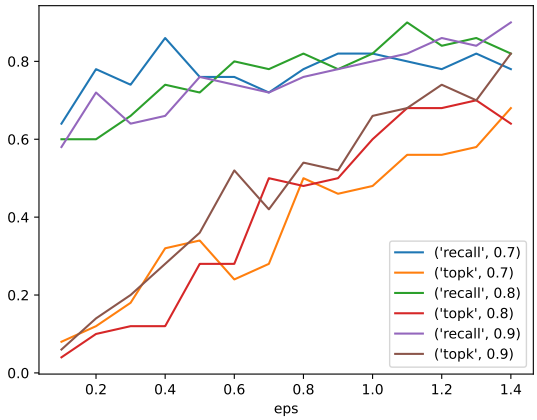

図 7　プライバシー費用 ϵ についての安全性指標 L-Mark の変化

表 13 にサンプルプログラムにかけた時の安全性指標の値の例を示す．図 7 には，一意率 0.5 を下回る様に第一匿名化したデータ C に対して，離散値列にランダマイズドレスポンス $p = 0.7, 0.8, 0.9$，連続値列にプライバシー費用 $\epsilon = 0.1, \ldots, 1.4$ のラプラスノイズを加えた時の，安全性指標の変化を示す．ϵ を下げる（ノイズを大きくする）につれて，所属推定 (recall, prec) もレコード識別率 (topk) の両方とも減少している．数値例を表 14 に示す．ϵ を大きく（ノイズを小さく）するにつれて，誤差 (rate, OR, coe) が小さく（緑）になるが，安全性（recall, prec, topk) が大きく（赤）くなっている．

表 14　有用性指標と安全性指標の関係

eps	rate	OR	cor	recall	prec	topk
0.1	0.138	0.521	0.250	0.613	0.609	0.067
0.2	0.080	0.456	0.207	0.667	0.662	0.113
0.3	0.069	0.401	0.204	0.693	0.662	0.153
0.4	0.068	0.311	0.220	0.700	0.695	0.273
0.5	0.068	0.314	0.218	0.727	0.722	0.327
0.6	0.067	0.421	0.223	0.800	0.785	0.413
0.7	0.068	0.358	0.205	0.773	0.758	0.420
0.8	0.067	0.483	0.211	0.767	0.762	0.493
0.9	0.067	0.369	0.212	0.800	0.789	0.533
1	0.067	0.402	0.202	0.793	0.788	0.527
1.1	0.071	0.333	0.210	0.813	0.808	0.607
1.2	0.066	0.378	0.218	0.847	0.847	0.633
1.3	0.070	0.305	0.202	0.833	0.833	0.740
1.4	0.069	0.375	0.224	0.860	0.860	0.720

3.7　加工サンプル

- Top coding

  `top2.py` 入力　列リスト　しきい値リスト 行番号

  例）`top2.py B.csv 1_5 75_50 e-top.csv`

  列 col でしきい値 theta より大きい行を出力する．例は，1 列 (age) が 75 歳以上，または，5 列 (bmi) が 50 以上 の行を出力する．
- Bottom coding

  `bottom2.py` 入力 列リスト しきい値リスト 行番号

  例）`bottom2.py B.csv 1_5 22_20 e-bot.csv`

  列でしきい値より小さい行を出力する．例は，1 列 (age) が 22 歳以下，5 列 (bmi) が 20 以下の行を出力する．
- k-anonimity

  `kanony2.py` 入力 `k`　列リスト 行番号

  例）`kanony2.py B.csv 7 2_3_4 e-ka.csv`

  列 columns を準識別子とみなして k-匿名を満たさない行を出力．例は，2 列 (人種), 3 列 (学歴)，4 列 (既婚歴) を準識別子とみなして，$k = 7$ で k-匿名性を満たさない行の番号を `e-ka.csv` に出力．
- Randomized Response

  `rr.py C p D`　列リスト

  例）`rr.py C.csv 0.9 d-xrr.csv 0_2_3_4_6_7_10`

  入力 C の中の指定された列を維持確率 p でランダマイズレスポンス（$1-p$ の確率で値域の中から一様な確率で置換える．例は，0,2,3,4,6,7,10 列を全て 0.9 の確率でランダマイズして，`d-xrr.csv` に出力している．
- Differential Privacy

  `dp2.py` 入力 列リスト ε リスト 出力

  例）`dp2.py d-xrr.csv 1_5 1.0_2.0 d-xrrdp.csv`

  差分プライバシーに基づいて指定された列に ϵ のラプラスノイズ $\frac{\epsilon}{2}e^{-\epsilon|x|}$ を加える．例は，1 列 (age) には $\varepsilon = 1.0$, 5 列 (bmi) には $\epsilon = 2.0$ のノイズを加えている．sensitivity $= 1$ とみなしているので，ϵ で調整する．

4.　個人情報の保護に対する配慮

本コンテストで用いるデータ “NHANES 2015-2016” は，NCHS Research Ethics Review Board (ERB) によって承認されている．コンテスト参加者は，米国 CDC の Web サイトから直接ダウンロードし，CDC の利用規定に従い自分で責任でデータの分析を行なう．PWS Cup 2021 コンテストにおいて，参加者は NHANES から生成された匿名化データを提出し，互いの匿名化データから出来るだけ多くのレコードを再識別することを競う．NHANES は匿名加工情報ではないが，個人情報保護法 [4] 38 条 (識別行為の禁止) に抵触する潜在的な可能性を避けるため，我々は他の参加者から提供されたデータはコンテスト参加者間でのみ共有し，外部へは開示しないことを奨励する．

5.　おわりに

匿名化コンテスト PWS Cup 2021 の趣旨，コンテストルール，糖尿病の罹患リスクを正確に，かつ，有用性高く加工するアルゴリズムを評価する指標を提案した．

謝辞

NHANES データについて有益なご助言を頂いた愛媛大学大学院 木村映善教授に感謝する．

参考文献

[1] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, (2021, `https://www.cdc.gov/nchs/nhanes/about_nhanes.htm`).

[2] Zhao, Fanfan, Wu, Wentao, Feng, Xiaojie, Li, Chengzhuo, Han, Didi, Guo, Xiaojuan, et al. (2020): Physical Activity Levels and Diabetes Prevalence in US Adults: Findings from NHANES 2015-2016. Adis Journals.

[3] “Suggested MET Scores”, Appendix 1, NHANES `https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/PAQ_I.htm#Appendix_1.__Suggested_MET_Scores`.

[4] 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 号及び平成 28 年法律第 51 号，令和 2 年法法律第 44 号により改正）, 2020. Personal Information Protection Commission, Japan, “Amended Act on the Protection of Personal Information”, June 2020. `https://www.ppc.go.jp/files/pdf/APPI_english.pdf`

[5] Josep Domingo-Ferrer, Vicenc Torra, “Information Loss: Evaulation and Measures”, *Data Privacy: Foundations, New Developments and the Big Data Challenge*, Springer, pp. 239 - 253, 2017.