PWS Cup 2021 ルール

予備戦 8/23版

PWS Cup WG

匿名化ヘルスケアデータコンテスト PWS Cup 2021

- ヘルスケアデータからの糖尿病リスクの評価
  - 性別，年齢，人種，学歴，BMIなどから，糖尿病(diabetes)に罹患するリスクを算出する.
  - 誰のレコードかわからない様に匿名化しても，リスクが正確に算出できることを目指す

Male	62	White	Gradschool	27.8	Not	1	Diabets
Male	53	White	HighSchool	30.8	Not	1	No
Male	78	White	HighSchool	28.8	Not	3	No
Female	56	White	Gradschool	42.4	Depressed	2	No
Female	42	Black	Colleage	20.3	Depressed	4	No
Female	72	Mexican	Gradschool	28.6	Not	1	No
Male	22	Black	Colleage	28	Not	1	No
Female	32	Mexican	Colleage	28.2	Not	1	No
Male	56	Black	HighSchool	33.6	Not	3	Diabets
Male	46	White	Gradschool	27.6	Not	2	No

ストーリー

- 登場人物

- 健康保険組合（＝加工者）
  - » 被保険者の検診データと治療データを持つ

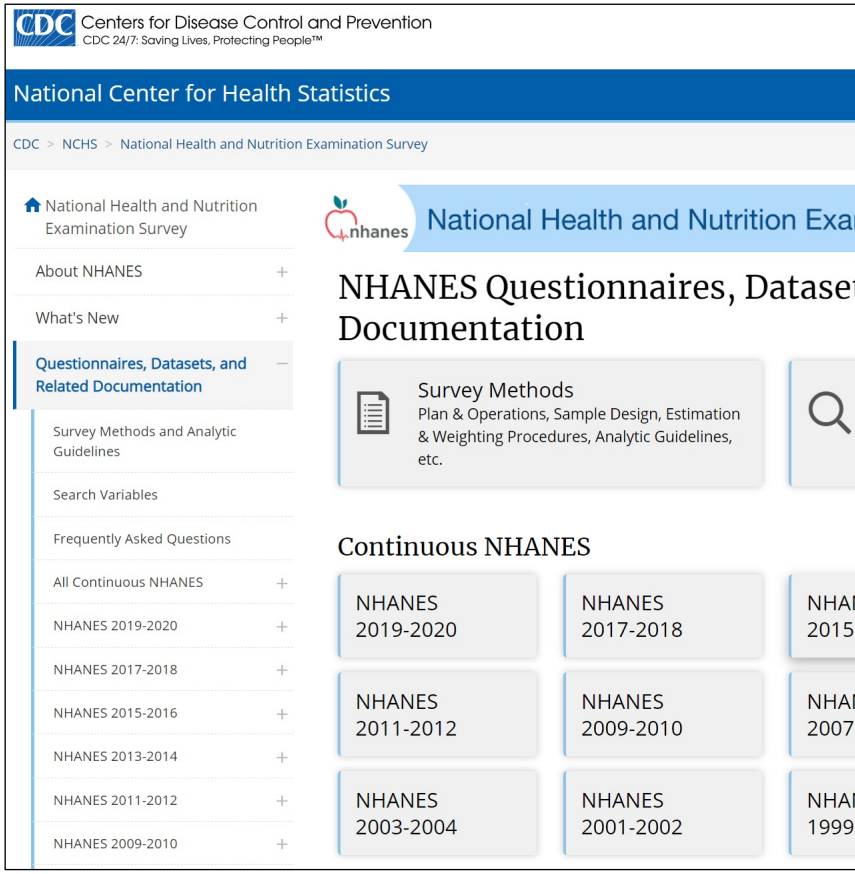
- 被保険者（＝攻撃者）
  - » 自分のデータが流通していないか心配・疑う

- 活用者（データ消費者）
  - » 匿名化されたデータから糖尿病罹患リスクを算出したい

- 審判（＝事務局）
  - » どの加工者が正しく安全に加工しているか判定

NHANES概要

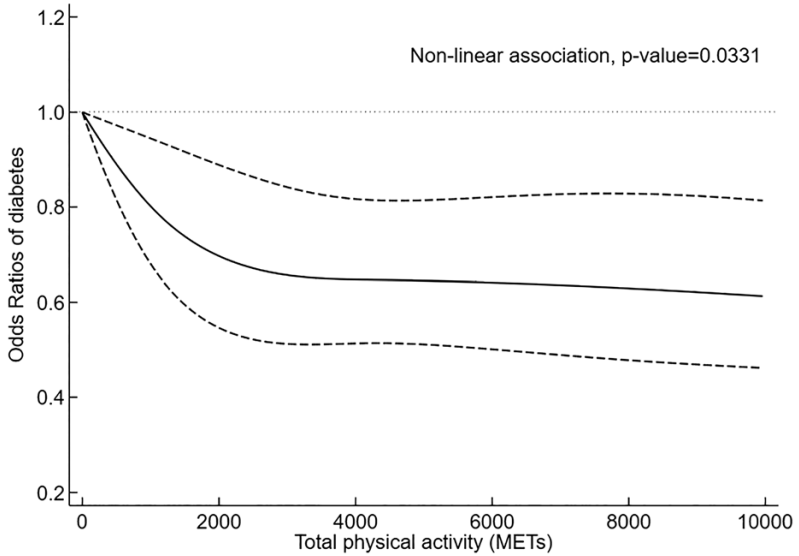
- National Health and Nutrition Examination Survey
  - **CDC** (米国疾病対策センター)の国民健康栄養調査プログラム
  - 1960年代から行われている調査．全米15箇所で，年5,000人を調査している．
  - 疫学研究，健全な公共健康政策やサービスの施策に活用．
  - 被験者世帯は，NCHS所長からのレターを受け取る．報酬と診断結果を得る．プライバシーは法律で守られている (privacy is protected by public laws)



利活用例「糖尿病罹患リスク」

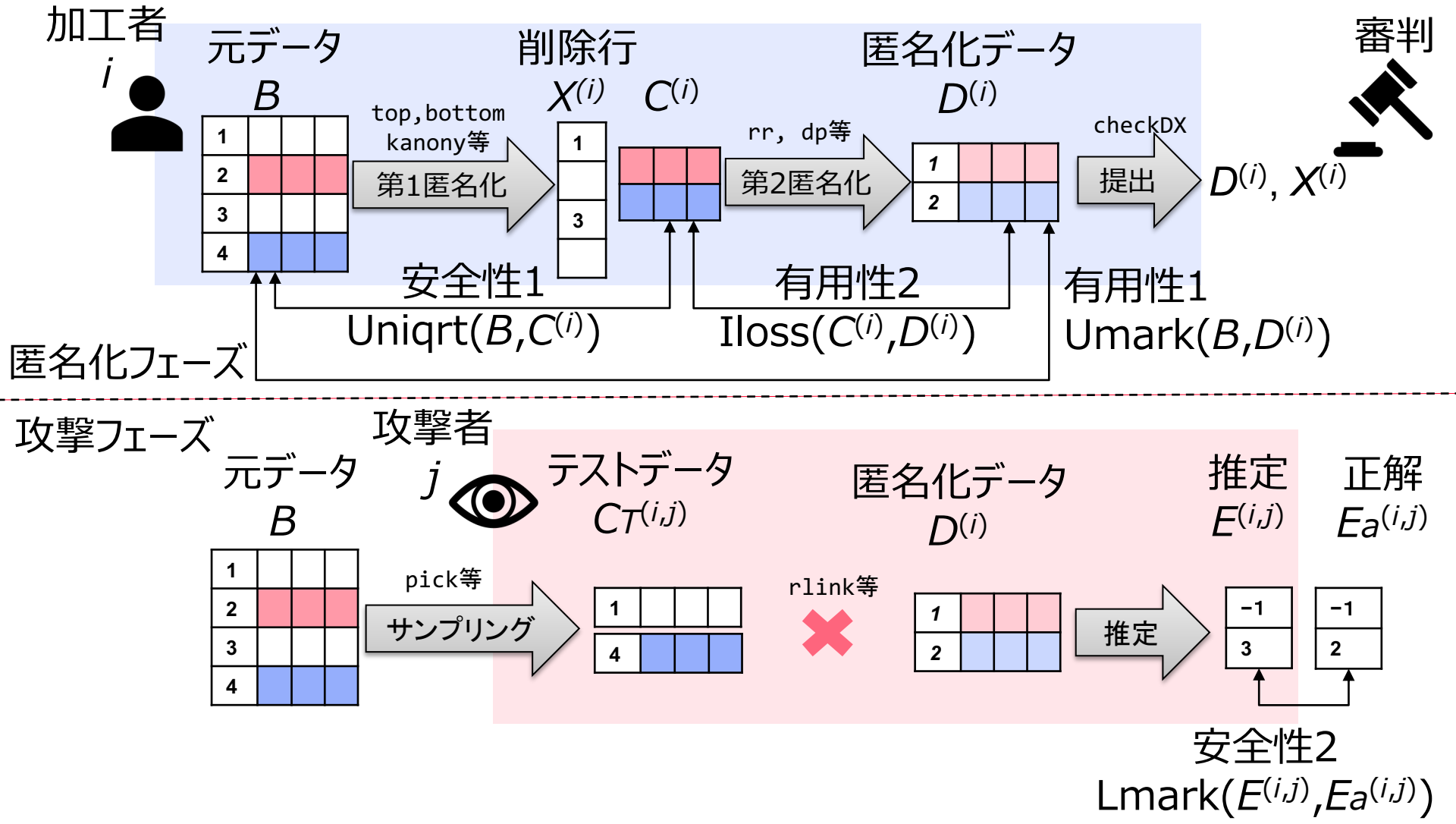
- ORs (Model III, 因子調整済み)

活動量 METs	オッズ比 OR	95% 信頼区間	p値
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



Fanfan Zhao, et. al (The First Affiliated Hospital of Jinan, Xi'an Jiaoton Univ.), "Physical Activity Levels and Diabetes Prevalence in US Adults: Findings from NHANES 2015–2016", Diabets Ther 2020.

ゲーム概要


加工者
i
元データ
B
削除行
X(i)
C(i)
匿名化データ
D(i)
審判
top,bottom
kanony等
第1匿名化
rr, dp等
第2匿名化
checkDX
提出
D(i), X(i)
安全性1
Uniqrt(B,C(i))
有用性2
Iloss(C(i),D(i))
有用性1
Umark(B,D(i))
匿名化フェーズ
攻撃フェーズ
攻撃者
j
元データ
B
テストデータ
CT(i,j)
匿名化データ
D(i)
推定
E(i,j)
正解
Ea(i,j)
pick等
サンプリング
rlink等
推定
安全性2
Lmark(E(i,j),Ea(i,j))

1. データの入手

test-1setup.sh

- CDCからデータを直接ダウンロードする
  - test-0config.sh の設定（チーム番号など）を修正
  - sh test-1setup.sh （Csv/B.csv が生成）

```
activ_diabet9_csv.py Csv/B.csv
```

12 列

gen	age	race	edu	mar	bmi	dep	pir	gh	mets	qm	dia
Male	62	White	Graduate	Married	27.8	0	0	0	0	Q2	1
Male	53	White	HighSchool	Divorced	30.8	0	1	0	0	Q1	0
Male	78	White	HighSchool	Married	28.8	0	0	0	0	Q3	1
Female	56	White	Graduate	Parther	42.4	1	0	0	0	Q3	0
Female	42	Black	College	Divorced	20.3	1	0	0	0	Q4	0
Female	72	Mexican	11th	Separated	28.6	0	0	0	0	Q1	0

n = 4,190 行

連続値
(第1列, 第5列)

未使用
(第8,9列)

目的変数
（糖尿病罹患）

2. 第1匿名化

test-2anonymize.sh

- 特異なデータを突き止め削除する
  - 条件
    - » (1) 行削除のみ. (値の加工はしない)
    - » (2) |B|/2 ≤ |C|　(半分を超えて削除しない)
    - » (3) uniqrt(C) ≤ 0.5 (Cの一意な行数の割合を0.5以下に)
    - » 例)22≤age ≤ 75, 20≤bmi≤50, k = 7で匿名化

```
top2.py Csv/dia6.csv 1_5 75_50 Csv/e-top.csv
bottom2.py Csv/dia6.csv 1_5 22_20 Csv/e-bot.csv
kanony2.py Csv/dia6.csv 7 2_3_4 Csv/e-ka.csv
join.py Csv/e-top.csv Csv/e-bot.csv $Csv/e-ka.csv > X.csv
exclude.py Csv/dia6.csv Csv/X.csv Csv/d-x.csv
umark.py Csv/dia6.csv Csv/d-x.csv

              cnt      rate      Coef        OR    pvalue       cor
max   417.000000  0.057100  0.629312  0.227358  0.346103  0.161865
mean   94.772727  0.007843  0.087464  0.059314  0.092200  0.006215
uniqrt.py Csv/C.csv
2360 0.7038473009245452 0.5632458233890215
```

3. 第2匿名化

test-2anonymize.sh

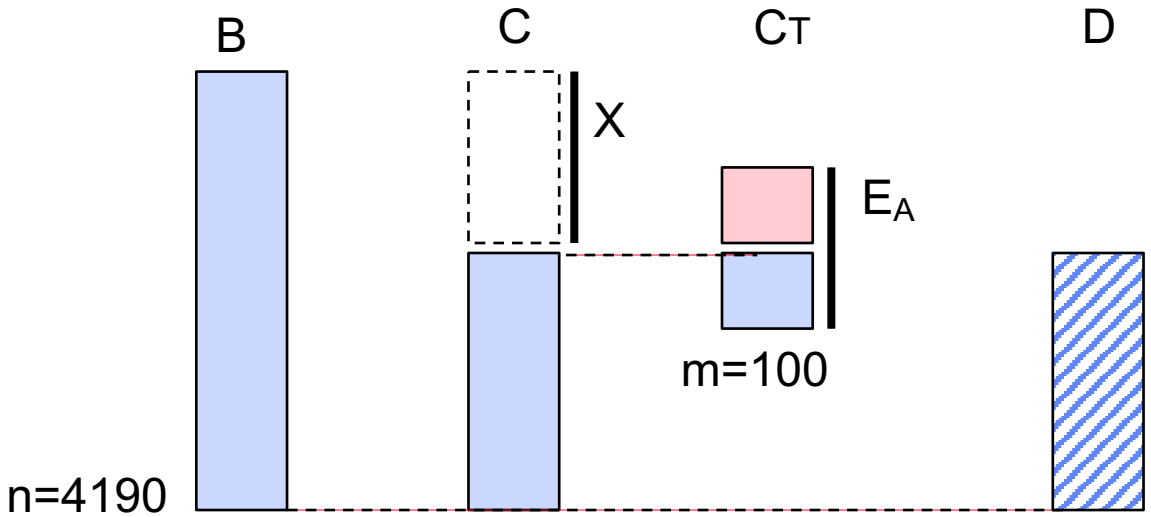
- 再識別させない様にデータを加工する
  - » (1) 値の変更のみ.（行は削除しない）
  - » (2) 離散値はBの値以外を取らない.
  - » 連続値は，Bの値域の範囲内　(13 ≤age≤ 85, 13 ≤bmi≤75)
  - » (3) 有用性指標の条件を満たす
    Umark(B,D): rate ≤ 0.05, OR ≤ 0.1, cor ≤ 0.1,
    Iloss(C,D):　iloss ≤ 6

```
rr.py Csv/d-x.csv 0.9 Csv/d-xrr.csv 0_2_3_4_6_7_10
dp2.py Csv/d-xrr.csv 1_5 1.0_2.0 Csv/d-xrrdp.csv
umark.py Csv/dia6.csv Csv/d-xrrdp.csv
              cnt      rate      Coef        OR    pvalue       cor
max    472.000000  0.063340  0.977877  0.676358  0.764306  0.173758
mean    95.030303  0.008682  0.137035  0.108739  0.175558  0.008086
iloss.py Csv2/C.csv Csv2/pre_anony_00_d.csv
             1         5       cat       max
mean  1.085297  0.723084  0.443483  1.085297
max   8.000000  4.400000  4.000000  8.000000
```

4. テストデータ生成(事務局) test-3pick.sh

審判

- 評価用データをサンプリングする
  - pick.py B.csv X.csv C.csv E_a.csv
    - » 入力: 元データB, 排除行X
    - » 出力: テストデータCT, 正解行番号 Ea
    - » Xから50(負例), B-Xから50(正例)をサンプリングする



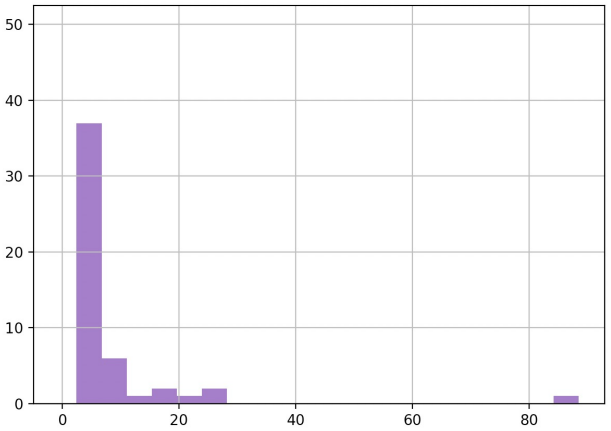
5. 攻撃

test-4rlink.sh

攻撃者

- 他チームの加工データiを推測する
  - 入力: テストデータ $C_T^{(i,j)}$, 匿名化データ$D^{(i)}$
  - 出力: 推測行番号 $E^{(i,j)}$
  - 例) rilink.py ユークリッド距離のメジアンについて，Cの各行がDに属するか（正）推定，推定行番号を上位3つEに推定

C_T	D1位ID	D2位ID	D3位ID	D1位距離	D2位距離	D3位距離	E(正解)
0	29	847	2599	0.0	5.3	120.0	29
1	-1	-1	-1	5.5	5.6	5.7	-1
2	2038	2345	2336	4.9	6.5	6.7	-1
3	2702	1378	2331	4.6	11.1	14.8	-1
4	134	1820	2580	0.0	3.1	3.7	134
5	138	967	2636	0.0	31.8	49.5	138
6	673	628	735	3.7	5.0	5.5	-1
7	189	1998	2268	0.0	15.6	24.2	189
8	-1	-1	-1	5.9	6.0	6.2	-1
9	302	2509	1973	0.0	5.7	6.5	302



距離の分布例

Lmark (Link benchmark) 安全性指標

- 推測の正しさの評価
  - 正解行番号Eaと推測行番号Eについて,
  - 所属推定:

$$recall = \frac{|{E_a}^X \cap E^X|}{|{E_a}^X|}, prec = \frac{|{E_a}^X \cap E^X|}{|E^X|}$$

  再識別率:

$$top_k = \frac{|\{\ell \in {E_a}^X | e_{a\ell} \in E_\ell\}}{|E_a^X|}$$

  - 例)
    - » E^X = {1,3,4,5}, Ea^X = {1,3,4}
    - » recall = 3/3 = 1.0
    - » prec = 3/4 = 0.75
    - » topk = 2/3 = 0.6

	$E_{\ell,1}$	$E_{\ell,2}$	$E_{\ell,3}$	E_a(正解)	判定
1	**29**	847	2599	29	✓
2	-1	-1	-1	-1	✓
3	2038	**2345**	2336	2345	✓
4	2702	1378	2331	80	NG
5	134	1820	2580	-1	NG

評価方法

審判

- 攻撃の総合評価
  - 危険度 = recall x prec x topk
- 匿名加工部門
  - 加工者iに対する推定の危険度の**最大**値（の低さ）で評価
- 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
- 予備戦1: 本戦9 の比で評価する.

Umark (**U**tility bench**mark**)有用性指標

rate
罹患率クロス集計
(33x2)

	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
Parther	331	57	0.079	0.014
Separated	101	32	0.024	0.008
Widowed	222	92	0.053	0.022
(15.0, 18.5]	64	4	0.015	0.001
(18.5, 25.0]	1027	114	0.245	0.027
(25.0, 30.0]	1149	241	0.274	0.058
(30.0, 70.0]	1075	514	0.257	0.123
dep_0	2666	647	0.636	0.154
dep_1	651	226	0.155	0.054
pir_0	2656	650	0.634	0.155

cor
共分散行列
(30x30)

	0_Male	0_Female	1	2_White	2_Black	2
0_Male	0.00	0.00	0.00	0.00	0.00	
0_Female	-1.00	0.00	0.00	0.00	0.00	
1	0.01	-0.01	0.00	0.00	0.00	
2_White	0.02	-0.02	0.12	0.00	0.00	
2_Black	-0.02	0.02	-0.05	-0.37	0.00	
2_Mexican	-0.02	0.02	-0.03	-0.34	-0.23	
2_Other	0.04	-0.04	-0.11	-0.30	-0.20	
2_Hispanic	-0.03	0.03	0.02	-0.29	-0.19	
3_Graduate	0.01	-0.01	-0.06	0.09	-0.07	
3_HighSchool	0.03	-0.03	-0.01	0.02	0.07	
3_College	-0.06	0.06	-0.05	0.11	0.07	
3_11th	0.04	-0.04	0.01	-0.10	0.02	
3_9th	-0.01	0.01	0.16	-0.21	-0.10	
3_nan	-0.02	0.02	0.02	-0.01	-0.01	
4_Married	0.10	-0.10	0.14	0.04	-0.15	
4_Divorced	-0.06	0.06	0.14	0.03	0.04	
4_Parther	0.02	-0.02	-0.21	-0.03	0.00	
4_Separated	-0.03	0.03	0.02	-0.06	0.04	
4_Never	0.00	0.00	-0.37	-0.08	0.16	
4_Widowed	-0.14	0.14	0.34	0.07	-0.01	
5	-0.08	0.08	0.06	-0.04	0.09	
6	-0.08	0.08	-0.03	0.03	-0.04	
7	-0.03	0.03	0.00	-0.21	0.04	
10_Q2	-0.06	0.06	0.09	0.00	0.01	
10_Q1	-0.11	0.11	0.21	-0.08	-0.03	
10_Q3	0.02	-0.02	-0.07	0.07	-0.04	

OR
オッズ比
(21x2)

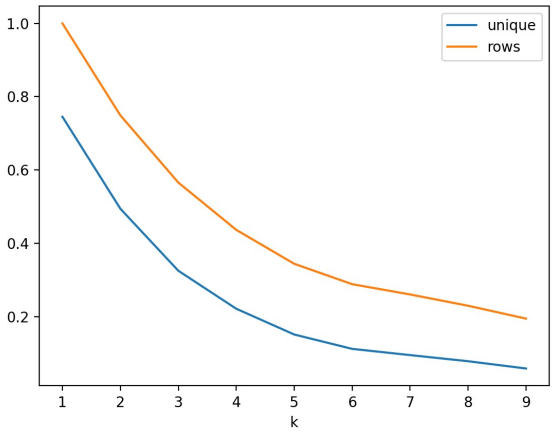
	Coef	OR	pvalue
Intercept	-7.319	0.001	0.0
gen[T.Male]	0.380	1.463	0.0
race[T.Hispanic]	-0.302	0.740	0.0
race[T.Mexican]	0.084	1.088	0.5
race[T.Other]	0.020	1.020	0.9
race[T.White]	-0.844	0.430	0.0
edu[T.9th]	-0.151	0.860	0.4
edu[T.College]	-0.073	0.930	0.6
edu[T.Graduate]	-0.150	0.861	0.3
edu[T.HighSchool]	-0.192	0.825	0.2
mar[T.Married]	0.278	1.320	0.0
mar[T.Never]	0.326	1.386	0.1
mar[T.Parther]	0.316	1.372	0.1
mar[T.Separated]	0.081	1.084	0.7
mar[T.Widowed]	-0.134	0.875	0.5
qm[T.Q2]	-0.228	0.796	0.0
qm[T.Q3]	-0.328	0.720	0.0
qm[T.Q4]	-0.401	0.670	0.0
age	0.057	1.058	0.0
bmi	0.097	1.102	0.0
dep	0.440	1.552	0.0
pir	0.147	1.158	0.1

Uniqrt (**Uniq**ue **rat**e) 安全性指標

- 定義
  - age (最小 20, 最大80) を10の位で丸める q=7値に
  - bmi (最小 15.5, 最大67.3)は，q=6値に
    - » 参考）race 5値，edu 5値，mar 6値，qm 4値
  - unique 率 = (Cの)一意な行数/(Bの)全行数

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

- 「一意」
  - その値の組を取る行が他にないこと(k = 1)
  - = attack.py で再識別される行数



lloss (Informtion Loss) 有用性指標

- lloss (セル間の情報損失量)
  - lloss(C,D) = max $_i$ d(C_i,D_i)
  - d(**x**,**y**) = 行**x**と行**y**のL1距離
    = $|x_i - y_i|$ if x_i が連続値(1,5列)
    = $|\{x_i \neq y_i\}|$ if x_i が離散値
  - 例）iloss(C,D) = max(9, 3, 5) = 9

C	Male	62	White	Graduate	Married	27.8	0	0	0	0	Q2	1
	0	9	0	1	1	3	0	1	0	0	1	1
D	Male	53	White	HighSchool	Divorced	30.8	0	1	0	0	Q1	0

	1	5	cat	max
mean	9	3	5	9
max	9	3	5	**9**

フォーマットのチェック

test-5check.sh

- 加工者
  - 加工 D，削除行 X

```
checkDX.py B.csv D.csv X.csv
D: num OK
D: obj OK
0 OK
2 OK
3 OK
4 OK
10 OK
(2724, 12) OK
X: int OK
X: unique OK
(695, 1) OK
```

- 攻撃者
  - checkE.py E

```
checkE.py B.csv E.csv
(100, 3) OK
E: int OK
E: max OK
E: min OK
```

全ての検査に OK が出るか，
提出前にそれぞれでチェックしてください．

お願い

- 他チームが匿名化したデータは二次配布しないで，参加者内でのみ共有ください．
  - 匿名加工情報ではないが，個人情報保護法38条（識別行為の禁止）を考慮するため
  - NHANESは倫理承認されており，CDCの趣旨に沿った分析には，追加の承認は不要
- 他のチームとの結託[*1]は禁止します．
  - 一般的な「情報共有」は奨励する．
- チームの代表者はCSS2021に参加登録を行い，最終プレゼンテーションをお願いします．
- ルールやしきい値などは，コンテストの途中で変更するかもしれないことご了承ください．

*1 事務局から受取るテストデータを共有する．独自のアイデアを含むプログラムを他チームに提供する．協力して特定のチームを攻撃する等．

参考文献

- （参照論文）
  - 菊池 浩明, 荒井 ひろみ, 井口 誠, 小栗 秀暢, 黒政 敦史, 千田 浩司, 中川 裕志, 中村 優一, 西山 賢志郎, 野島 良, 村上 隆夫, 波多野 卓磨, 濱田 浩気, 古川 諒, 馬 瑞強, 前田 若菜, 山岡 裕司, 山田 明, 渡辺 知恵美, “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”, 情報処理学会コンピュータセキュリティシンポジウム 2021, プライバシーワークショップ (PWS), 2021. (発表予定）
- （NHANES出典, 引用形式）
  - 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 Hu- man Services, Centers for Disease Control and Prevention.

2021 PWS Cup WG

コンテストに参加しているメンバーもいます

菊池 浩明 *	明治大学総合数理学部
千田 浩司	NTTセキュアプラットフォーム研究所
野島 良	国立研究開発法人 情報通信研究機構 ネットワークセキュリティ研究所
黒政 敦史 *	富士通クラウドテクノロジーズ株式会社
濱田 浩気	NTTセキュアプラットフォーム研究所
荒井 ひろみ	理化学研究所
村上 隆夫	国立研究開発法人 産業技術総合研究所
山岡 裕司	富士通株式会社
小栗 秀暢	富士通株式会社
渡辺 知恵美	筑波技術大学
中川 裕志	理化学研究所
波多野 卓磨	日鉄ソリューションズ
山田 明	KDDI総合研究所
西山 賢志郎 *	株式会社ビズリーチ
中村 優一 *	早稲田大学
井口 誠	Kii 株式会社
古川 諒	NEC セキュリティシステム研究所
馬 瑞強	明治大学
前田 若菜	富士通株式会社

*事務局