
PWS Cup 2021 ルール

予備戦 8/19版

PWS Cup WG (菊池)

Andrew を探せ

■「特異な記述」は削除する



果たして Andrew のプライバシーは守られていると言えるのでしょうか

現実の匿名加工方法の課題

No	19条規則	加工方法	該当
1	特定の個人を識別出来る記述等の全部または一部を削除	削除, 仮名化(規則性のない方法で生成された仮IDに置換)	○
2	個人識別符号の全部を削除	削除	少
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	○
4	特異な記述等を削除	トップ・ボトムコーディング	○
5	他の個人情報との差異等の性質を勘案した適切な措置	k-匿名化	なし

第三者提供する匿名加工情報に含まれる個人に関する情報

- DPC データ(E ファイル・F ファイル・様式 1、D ファイル)
- E ファイル及び F ファイル形式の外来データ
- 電子レセプト(医科及び DPC)

氏名・住所・電話番号は含まれません。また、生年月日・郵便番号・各種保険証に関する情報は下記のように変換されます。

- ・ 生年月日: 生年月及び入院時年齢に変換
- ・ 郵便番号: 上 3 桁のみに変換(下 4 桁を削除)
- ・ 各種保険証に関する情報: 保険者番号(※)のみに変換

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)

The screenshot shows the official website for the National Health and Nutrition Examination Survey (NHANES) managed by the CDC. The header includes the CDC logo and the text 'Centers for Disease Control and Prevention' and 'CDC 24/7: Saving Lives, Protecting People™'. Below this is a blue navigation bar with 'National Center for Health Statistics'. The main content area has a breadcrumb trail: 'CDC > NCHS > National Health and Nutrition Examination Survey'. On the left is a sidebar menu with links: 'National Health and Nutrition Examination Survey', 'About NHANES', 'What's New', 'Questionnaires, Datasets, and Related Documentation' (which is expanded to show 'Survey Methods and Analytic Guidelines', 'Search Variables', and 'Frequently Asked Questions'), 'All Continuous NHANES', and a list of survey cycles from 2019-2020 down to 2009-2010. The main content area features the 'NHANES Questionnaires, Dataset, and Documentation' section with a 'Survey Methods' link. Below this is the 'Continuous NHANES' section, which displays a grid of survey cycles: 2019-2020, 2017-2018, 2011-2012, 2009-2010, 2003-2004, 2001-2002, and 1999.

CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

National Center for Health Statistics

CDC > NCHS > National Health and Nutrition Examination Survey

National Health and Nutrition Examination Survey

About NHANES +

What's New +

Questionnaires, Datasets, and Related Documentation -

Survey Methods and Analytic Guidelines

Search Variables

Frequently Asked Questions

All Continuous NHANES +

NHANES 2019-2020 +

NHANES 2017-2018 +

NHANES 2015-2016 +

NHANES 2013-2014 +

NHANES 2011-2012 +

NHANES 2009-2010 +

NHANES Questionnaires, Dataset, and Documentation

Survey Methods
Plan & Operations, Sample Design, Estimation & Weighting Procedures, Analytic Guidelines, etc.

Continuous NHANES

NHANES 2019-2020

NHANES 2017-2018

NHANES 2011-2012

NHANES 2009-2010

NHANES 2003-2004

NHANES 2001-2002

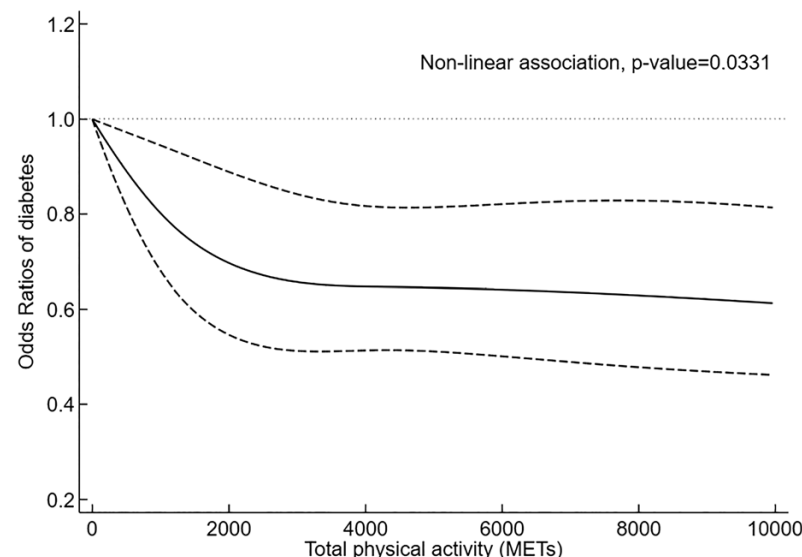
NHANES 1999

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	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 Jiaotong Univ.),
“Physical Activity Levels and Diabetes Prevalence in US Adults: Findings from
NHANES 2015–2016”, Diabetics Ther 2020.

ゲーム概要

加工者
 i



元データ
 B

1			
2			
3			
4			

top, bottom
kanony等
第1匿名化

削除行
 X_i

1
3

rr, dp等
第2匿名化

匿名化データ
 D_i

1			
2			

checkDX
提出

D_i, X_i

審判



安全性1

$\text{Uniqrt}(B, C_i)$

有用性2

$\text{Iloss}(C, D_i)$

有用性1

$\text{Umark}(B, D_i)$

攻撃者
 j



元データ
 B

1			
2			
3			
4			

pick 等

サンプリング

テストデータ
 CT_j

1			
4			

rlink 等



匿名化データ
 D_i

1			
2			

推定

推定
 E_j

-1
3

正解
 EA_i

-1
2

安全性2

$\text{Lmark}(E_j, EA_i)$

1. データの入手

- 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 test1setup.sh (Csv/B.csv が生成)
activ_diabet9_csv.py Csv/B.csv

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



2. 第1匿名化

■ 特異なデータを突き止め削除する

□ 条件

- » 1. 行削除のみ. $|B|/2 \leq |C|$ (半分を超えて削除しない)
- » 2. $\text{uniqrt}(C) < 0.5$ (Cの一意的な行数の割合が0.5未満にする)
- » 例) $22 \leq \text{age} \leq 75$, $20 \leq \text{bmi} \leq 50$, $k = 7$ 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



加工者

■ 再識別させない様にデータを加工する

- » 3. 値の変更のみ. (行は削除しない)
- » 4. 離散値はBの値以外を取らない.
- » 5. 連続値は, $13 \leq \text{age} \leq 85$, $13 \leq \text{bmi} \leq 75$ の範囲内
- » 6. Umark(B,D): $\text{rate} \leq 0.05$, $\text{OR} \leq 0.1$, $\text{cor} \leq 0.1$,
- » 7. lloss(C,D): $\text{iloss} \leq 6$

```
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



審判

4. テストデータ生成(事務局)

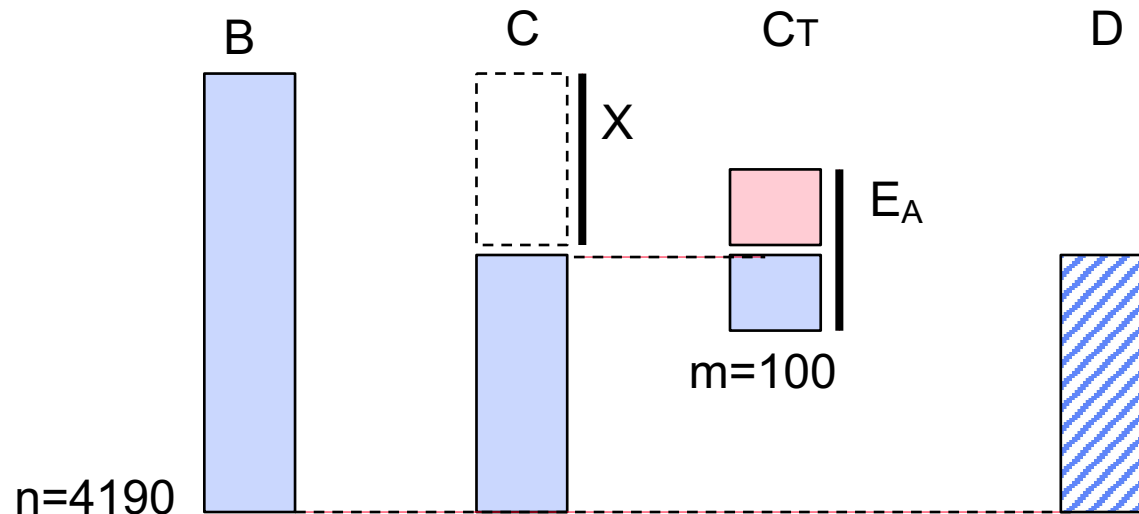
■ 評価用データをサンプリングする

□ pick.py B.csv X.csv C.csv E_a.csv

» 入力: 元データB, 排除行Ex

» 出力: テストデータC_T, 正解行番号 E_a

» Exから50(負例), B-Xから50(正例)をサンプリングする



5. 攻撃



攻撃者

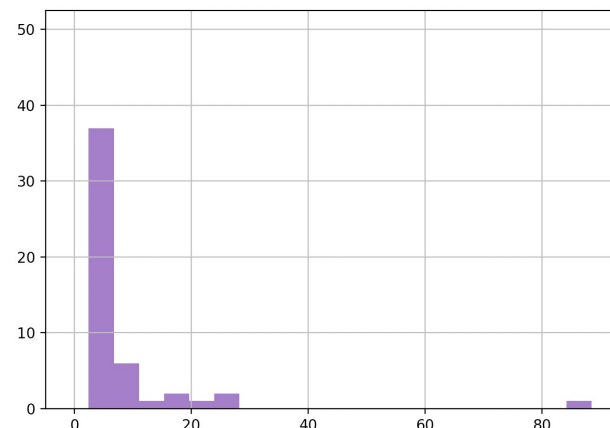
■ rlink.py C_T.csv D.csv E.csv

□ 入力: テストデータ C_T, 匿名化データD

□ 出力: 推測行番号 E

□ 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	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
10	309	1148	7	0.0	15.2	16.5	309



Lmark (Link benchmark) 安全性指標



審判

■ lmark.py Ea.csv E.csv

□ 正解行番号Eaと推測行番号Eについて,

□ メンバーシップ推定:

$$\text{recall} = |Ea^X \cap E^X| / |Ea^X|,$$

$$\text{precision} = |Ea^X \cap E^X| / |E^X|$$

□ 再識別率(top-k): $|\{\ell \in Ea^X \mid Ea_\ell \in E_\ell\}| / |Ea^X|$

□ 例)

» $\text{recall} = 3/3 = 1.0$

» $\text{prec} = 3/4 = 0.75$

» $\text{topk} = 2/3 = 0.6$

D1位ID	D2位ID	D3位ID	E _A (正解)	topk
29	847	2599	29	✓
-1	-1	-1	-1	
2038	2345	2336	2345	✓
2702	1378	2331	80	NG
134	1820	2580	-1	NG

評価方法

- 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 (Utility benchmark)有用性指標

rate
罹患率クロス集計
(33x2)

	cnt		rate	
	0	1	0	1
diabetes				
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	0.00
gen[T.Male]	0.380	1.463	0.00
race[T.Hispanic]	-0.302	0.740	0.00
race[T.Mexican]	0.084	1.088	0.51
race[T.Other]	0.020	1.020	0.90
race[T.White]	-0.844	0.430	0.00
edu[T.9th]	-0.151	0.860	0.40
edu[T.College]	-0.073	0.930	0.63
edu[T.Graduate]	-0.150	0.861	0.35
edu[T.HighSchool]	-0.192	0.825	0.24
mar[T.Married]	0.278	1.320	0.00
mar[T.Never]	0.326	1.386	0.10
mar[T.Parther]	0.316	1.372	0.10
mar[T.Separated]	0.081	1.084	0.71
mar[T.Widowed]	-0.134	0.875	0.51
qm[T.Q2]	-0.228	0.796	0.00
qm[T.Q3]	-0.328	0.720	0.00
qm[T.Q4]	-0.401	0.670	0.00
age	0.057	1.058	0.00
bmi	0.097	1.102	0.00
dep	0.440	1.552	0.00
pir	0.147	1.158	0.17

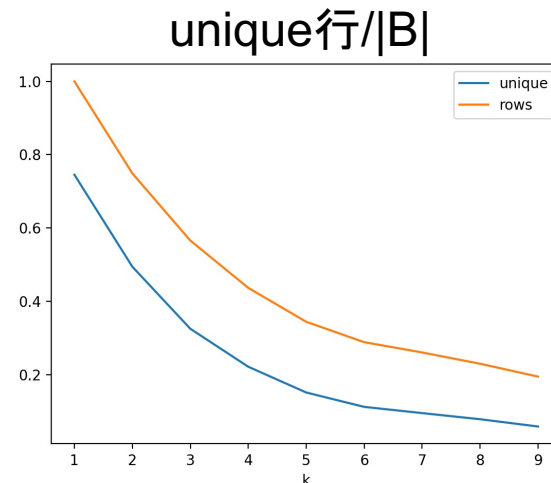
Uniqrt (Unique rate) 安全性指標

■ 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

■ 「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



iloss (Informtion Loss) 有用性指標

■ iloss (セル間の情報損失量)

□ $\text{iloss}(C,D) = \max_i d(C_i, D_i)$

□ $d(\mathbf{x}, \mathbf{y}) = \text{行}\mathbf{x}\text{と行}\mathbf{y}\text{のL1距離}$

$= |x_i - y_i|$ if x_i が連続値(1,5列)

$= |\{x_i \neq y_i\}|$ if x_i が離散値

□ 例) $\text{iloss}(C,D) = 9$

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

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。