


# PWS Cup 2022

COVID-19 重篤化患者のプライバシーを守り切れるか？

## ルール説明

PWS Cup ワーキンググループ グループ長 野島 良



# PWS Cup 2022

## 匿名ヘルスケアデータコンテスト

- COVID-19 重篤化患者のプライバシーを守り切れるか？
  - コロナ重篤化患者の、性別、年齢、人種、学歴、病歴などからなるデータ
    - [\*]において、NHANESをベースとした合成データが作成
    - 本合成データを参考
  - 誰のレコードかわからない様に匿名化しても、コロナで重篤化するリスクを算出できることを目指す

[\*]B. Seligman, M. Ferranna, D.E. Bloom, Social determinants of mortality from COVID-19: A simulation study using NHANES, PLOS Medicine 18(12): e1003888



# ストーリー

---

- 登場人物
  - 加工者：コロナ重篤化患者からなるデータを匿名化する
  - 攻撃者：匿名化されたデータから、知人が重篤化したかどうかを特定する
  - 活用者：匿名化されたデータから、重篤化リスクを算出する
  - 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



# データ

## 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)

Centers for Disease  
Control and Prevention

Center for Health Statistics

Health and Nutrition Examination Survey

National Health and Nutrition  
Examination Survey

Participants



Survey Data and



# スケジュール（予定）

---

5月

- データセット整備, ルール案
- ポスター, ウェブ

6月

- 有用性評価,
- 安全性評価

7月

- トライアル
- 参加者募集  
開始7/22～

8月

- 予備戦 匿名化  
2022/8/18-  
8/30
- 予備戦 攻撃  
2022/9/2-  
9/13

9月

- 本戦 匿名化  
2022/9/16-  
10/3
- 攻撃フェーズ  
2022/10/7-  
10/18

10月

- リハーサル
- CSS当日ポスターセッション  
(1日)

ファイルダウンロード

匿名化フェーズ

CORE

23

dmainA\_2\_0\_0.csv  
original\_data1\_0.txt  
2022.06.15 up

HIGH SCORE

0.873

スコア

Uploads 384

B	0.873
D	0.865
S	0.839
C	0.798
F	0.796
J	0.789
Q	0.782

チームスコア推移



元ファイルダウンロード

匿名化フェーズ

攻撃フ

TEAM FILE

dmainA\_2\_0\_0.csv

0.268  
[73]

dmainA\_2\_1\_0.csv

0.278  
[67]

dmainA\_2\_2\_0.csv

0.313  
[19]

dmainA\_2\_3\_0.csv

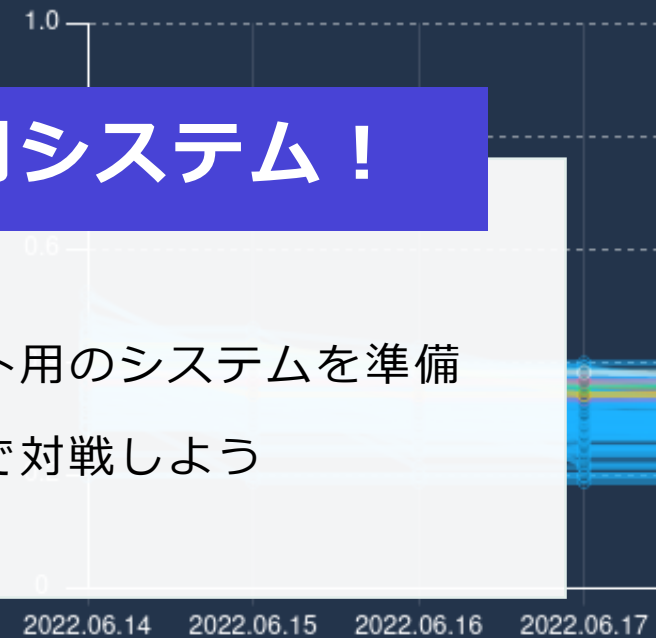
0.305  
[30]

匿名化ファイルランク

Uploads 99

1	O	dmainO_2_3_2.csv	0.378 [91]
2	G	dmainG_2_3_2.csv	0.378 [91]
3	D	dmainD_2_3_2.csv	0.378 [91]
4	E	dmainB_2_3_0.csv	0.356 [91]
5	E	dmainE_2_3_0.csv	0.356 [91]
6	D	dmainD_2_3_0.csv	0.356 [91]
7	E	dmainE_2_0_3.csv	0.336 [91]
8	O	dmainO_2_2_2.csv	0.335 [91]

チームスコア推移



コンテスト用システム！

- 今年度は、コンテスト用のシステムを準備
- みんなでオンラインで対戦しよう

# 参加方法とアクセス方法

- 参加方法（2022/7/22～）

PWSCUP2022ホームページ

<https://www.iwsec.org/pws/2022/cup22.html>

PWSCUP2022参加規程

<https://www.iwsec.org/pws/2022/entry.html>

をよくお読みになり、エントリーフォームから、お申し込みください。

- アクセス方法

PWSCUP2022事務局から参加登録完了メールが届きます。

その後、大会システムのURLとIDとパスワード、利用規約をメールでうけとり、大会のシステムのマイページにアクセス可能となります。

# コンテスト概要

- 加工者：重篤化した患者のデータ (D) を匿名化 (D') する
- 攻撃者：知人が重篤化したかどうかを知りたい

匿名化フェーズ

加工者

1			
2			
3			
4			
...			



1			
2			
3			
4			
...			

D: 加工前データ

D': 加工後データ

攻撃フェーズ

攻撃者

1			
2			
3			
4			
...			

D': 加工後データ



入っている・  
入っていない

--	--	--	--

r: 知人のデータ

メンバーシップ推定





スミレちゃん



元ファイルダウンロード

SCORE

0.623

dmainA\_2\_0\_0.csv  
original\_data1\_0.txt  
2022.06.15 up

コードリスト

Uploads 4

A\_2\_0\_0.csv

0.623

\_data1\_0.txt

5.15 up

A\_2\_1\_0.csv

0.546

\_data1\_0.txt

5.18 up

A\_2\_2\_0.csv

0.503

\_data1\_0.txt

ファイル管理

TEAM SCORE

0.623

dmainA\_2\_0\_0.csv  
original\_data1\_0.txt  
2022.06.15 up

全チームスコア

Uploads 384

1

TEAM B

0.873

2

TEAM D

0.865

3

TEAM S

0.839

4

TEAM C

0.798

5

TEAM F

0.796

6

TEAM J

0.789

7

TEAM Q

0.782

匿名化フェーズ

攻撃フェーズ

PWSCUP

## 匿名化フェーズ

1. 元データをダウンロード
2. 事務局が提供する6つの加工手法を使い、  
データを加工
  - **top2.py** (トップコーディング)
  - **bottom2.py** (ボトムコーディング)
  - **kanony2.py** (k-匿名)
  - **exclude.py** (行排除)
  - **rr.py** (ランダムイズ)
  - **dp2.py** (差分プライバシー)
3. 結果をアップロード
4. システム上で自動採点
5. 複数の結果の中から1つを選び提出

HIGH SCORE

0.873

チームスコア

1.0

0.8

0.6

0.4

0.2

0

2022.06.14

2022.06.15

2022.06.16

2022.06.17

2022.06.18

2022.06.19

2022.06.20

2022.06.21

# 匿名化フェーズ詳細（加工方法）

- 課題となるデータは6種類用意されており、マイページからダウンロードできます。
- ダウンロードしたデータは、事務局が提供する6つのアルゴリズム（Python3）プログラムを使って各自の環境で加工し、加工手法のログとともにマイページにアップロードし、数秒後、採点結果を確認できます。
- データ加工した結果は複数アップロード可能（条件付き）ですが、最終結果としてその中から1つ選んで提出してください。

注）加工に利用するアルゴリズムはPWSCUPシステムのダウンロードサイトから入手が可能です

提出した匿名化データは攻撃フェーズでほかのチームからの攻撃対象データとして使われます。

# 提出物

加工データアップロード時に

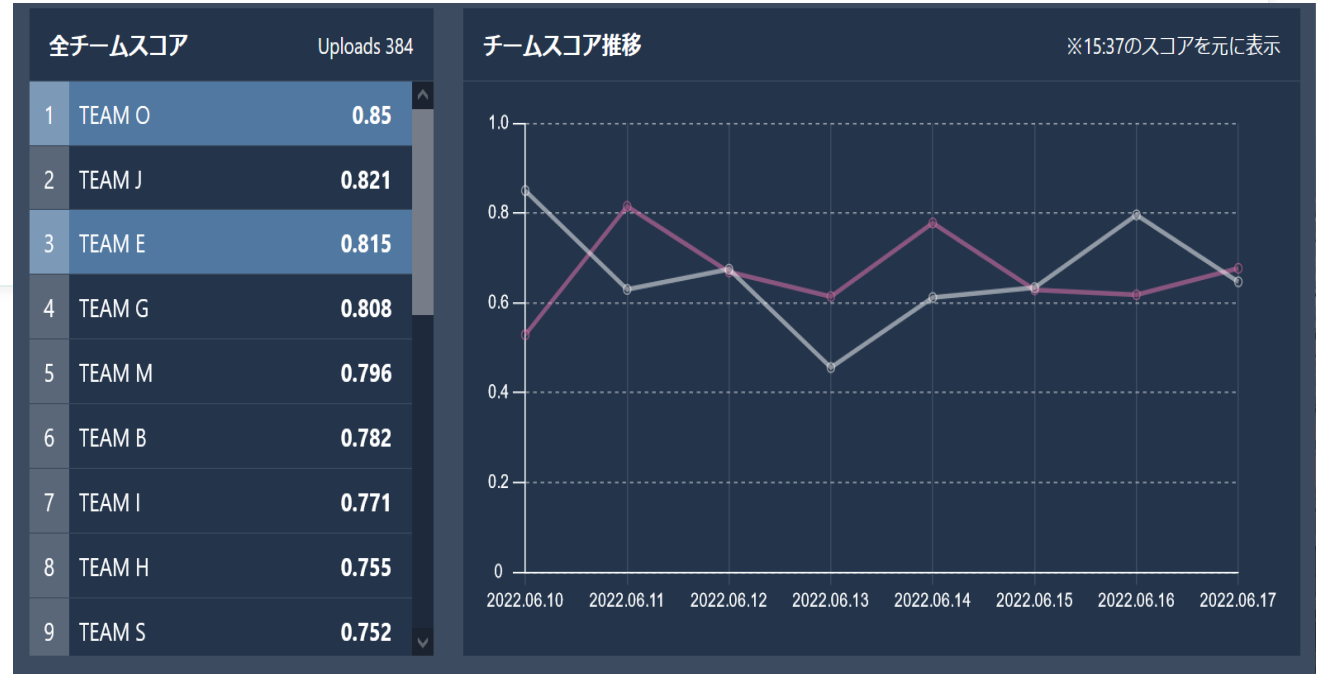
1. 加工データ
2. 加工に用いたアルゴリズムのログデータを提出します。

提出ファイルはアップロード時にフォーマットチェックが行われ、正しいデータがシステムに登録されます。

ログデータの例（正式なものは大会開始時に公開）：

```
python ../Anon/kanony2.py ../Data/orig_data1.csv anon_data1_k.csv 2 1_2
python ../Anon/rr.py anon_data1_k.csv anon_data2_rr.csv 0.2 1_2 31
python ../Anon/dp2.py anon_data2_rr.csv anon_data2_rr.csv 0 0.1 31
python ../Anon/top2.py anon_data2_rr.csv anon_data2_rr.csv 0 80
python ../Anon/bottom2.py anon_data2_rr.csv anon_data2_rr.csv 4 1
python ../Anon/shuffle.py anon_data2_rr.csv anon_data2_rr.csv 4
python ../Anon/exclude.py anon_data2_rr.csv anon_data2_rr.csv 4
```

例：これを提出する加工データとしてください。



YOUR SCORE

0.426

amainA\_3\_46.txt  
K dmainK\_2\_2\_2.csv  
2022.06.16 up

アップロードリスト

Uploads 95

mainA_3_7.txt	0
dmainC_2_0_0.csv	2022.06.14 up
mainA_3_8.txt	0
dmainC_2_1_4.csv	2022.06.21 up
mainA_3_26.txt	0
dmainG_2_3_3.csv	2022.06.14 up
mainA_3_22.txt	0.243
dmainF_2_3_0.csv	

ファイル管理

TEAM FILE

dmainA\_2\_0\_0.csv

0.268

[73]

dmainA\_2\_1\_0.csv

0.278

[67]

dmainA\_2\_2\_0.csv

0.31

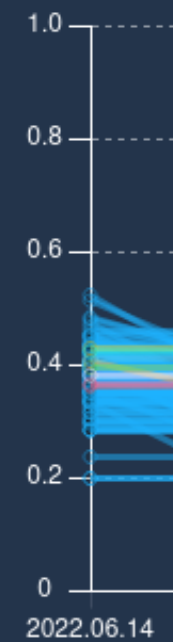
[19]

匿名化ファイルリンク

Uploads 99

1	<b>O</b>	dmainO_2_3_2.csv	0.378 [91]
2	<b>G</b>	dmainG_2_2_0.csv	0.367 [91]
3	<b>D</b>	dmainD_2_1_0.csv	0.363 [91]
4	<b>B</b>	dmainB_2_3_0.csv	0.356 [91]
5	<b>E</b>	dmainE_2_2_4.csv	0.356 [91]
6	<b>L</b>	dmainL_2_1_2.csv	0.343 [91]
7	<b>E</b>	dmainE_2_0_3.csv	0.336 [91]
8	<b>O</b>	dmainO_2_2_2.csv	0.335 [91]

チームスコア



## 攻撃

①他のチームが作成した加工データ

②事務局が用意した知人のデータ

を使って、

知人が「加工後のデータ」に含まれているかどうか

を推定します。

- 知人のリストは2X名から構成されています。
- X名がデータに含まれており、X名が含まれていない構成となっています。
- 的中させた人数をZとしたときに、 $(2X - Z)/2X$  を得点とします。
- 攻撃回数は平等性を期すため制限があります。

1. 攻撃対象データを選び、攻撃対象データと知人データを

**ダウンロード**

2. 推定データを作成

3. 結果を**アップロード**

4. システム上で**自動採点**



# 総合評価（予定）

## ・予備戦・本戦

- ・有用性評価：3つの評価手法の平均。それぞれ最低0、最大1
- ・安全性評価：最大1、最低0（破られていないものの割合）
- ・評価：有用性評価と安全性評価の平均

## ・総合評価

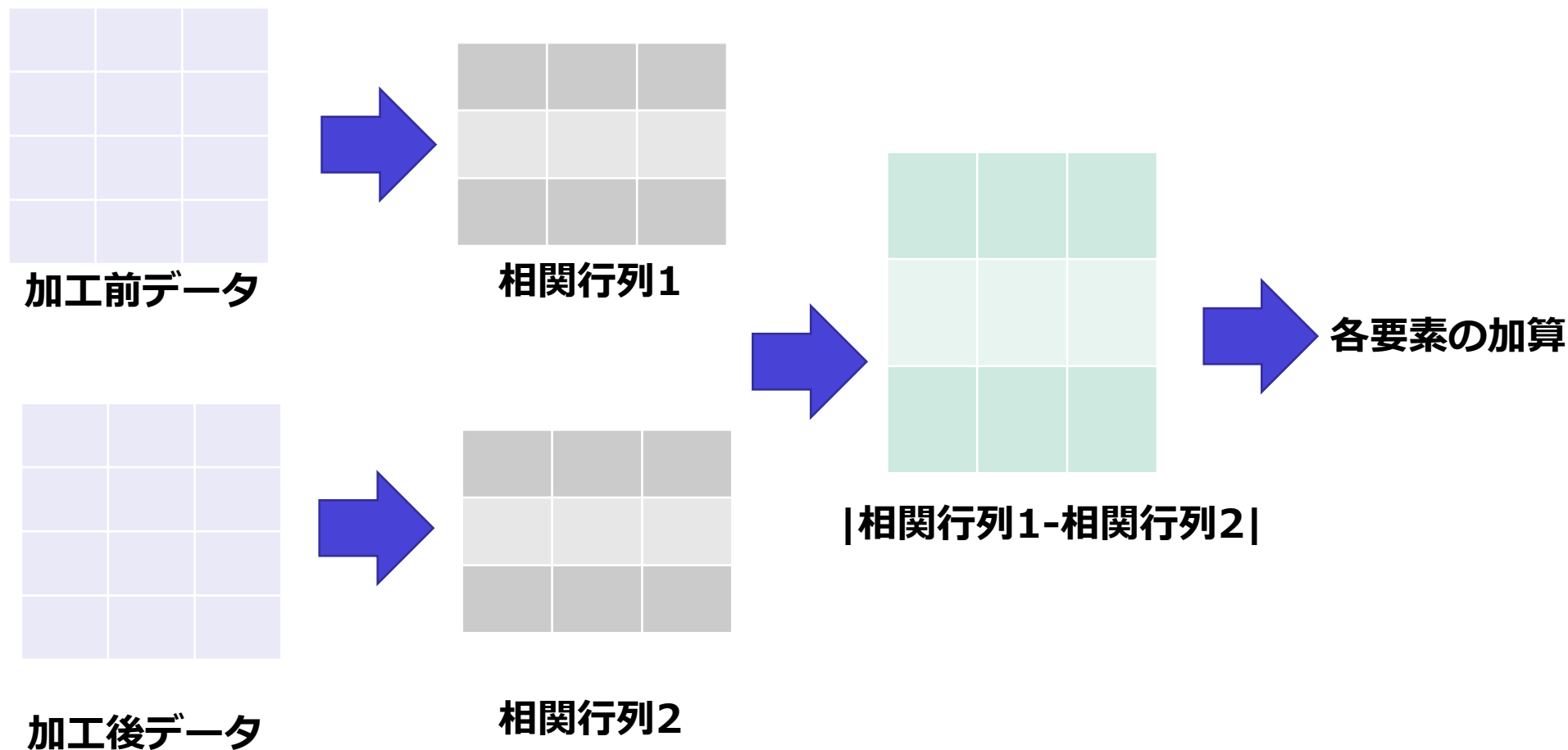
- ・予備戦 1 対 本戦 9 の比で評価



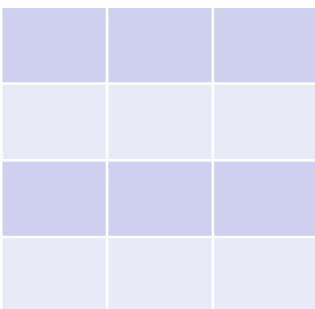
# 有用性評価 重篤化リスク評価

- ロジスティック回帰分析：
  - COVID ~ AGE + GENDER + RACE + INCOME + EDUCATION + VETERAN + NOH + HTN + DM + IHD + CKD + COPD + CA
- 加工データ、加工前データにおいてそれぞれロジスティック回帰分析を行う
- それぞれの偏回帰係数（の指数乗）の差分の合計(OR比)
- コード名

## 有用性評価2 相関行列の差（予定）



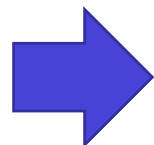
# 有用性評価 集計数の差（予定）



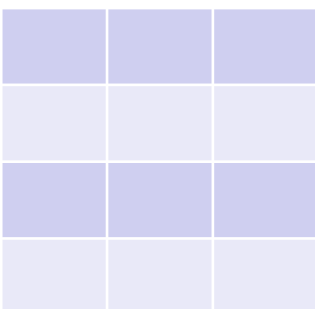
加工前データ



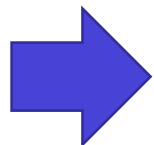
女性の数X1  
男性の数Y1



$$|X1 - X2| + |Y1 - Y2| + |Z1 - Z2|$$



加工後データ



女性の数X2  
男性の数Y2  
99の数 Z2

全ての属性において実行し、全てを合算する

# お願い

- チームの代表者はCSS2022に参加登録を行い，最終日にプレゼンテーションをお願いします．
- ルール・システムなど、まだ検討中で変更するかもしれないことをご了承ください．

（留意事項） NHANESは倫理承認されており，CDCの趣旨に沿った分析には，追加の承認は不要であることをご承知おきください