


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

5月	6月	7月	8月	9月	10月
<ul style="list-style-type: none">• データセット整備, ルール案• ポスター, ウェブ準備	<ul style="list-style-type: none">• 有用性評価• 安全性評価	<ul style="list-style-type: none">• トライアル• 参加者募集 開始7/22～	<ul style="list-style-type: none">• 予備戦 匿名化 2022/8/18-8/30• 予備戦 攻撃 2022/9/2-9/13	<ul style="list-style-type: none">• 本戦 匿名化 2022/9/16-10/3• 攻撃フェーズ 2022/10/7-10/18	<ul style="list-style-type: none">• 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

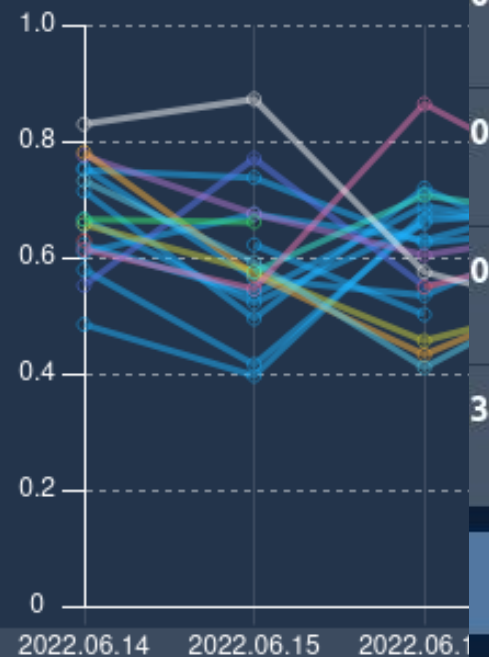
0.873

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

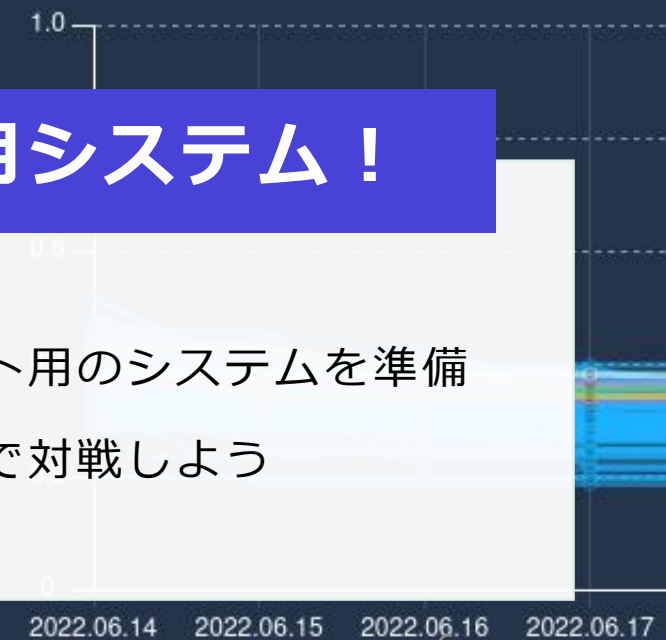
0.305
[30]

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	0.375 [91]
3	D	dmainD_2_3_2.csv	0.375 [91]
4	E	dmainB_2_3_0.csv	0.356 [91]
5	E	dmainA_2_3_2.csv	0.356 [91]
6	L	dmainL_2_3_2.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。

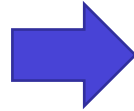
コンテスト概要

- 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

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

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

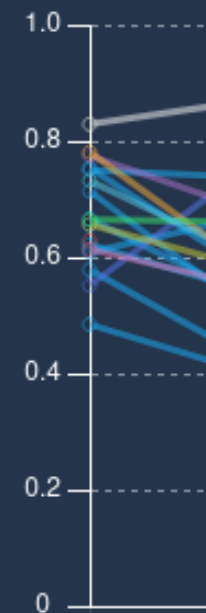
6 TEAM J 0.789

7 TEAM Q 0.782

HIGH SCORE

0.873

チームスコア



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

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。

匿名化に使うデータ（加工前データ）

次のようなデータを匿名化してもらいます

コロナ重篤化した人のデータセット (300名)

病歴

AGE	GENDER	RACE	INCOME	EDUCATION	VETERAN	NOH	HTN	DM	IHD	CKD	COPD	CA
66	2	4	3	2	0	1	1	0	0	0	0	0
80	2	4	4	3	0	2	1	0	0	0	0	0
76	1	4	7	4	1	1	0	0	0	0	0	0
50	1	6	15	5	0	3	1	0	0	0	0	0
77	2	3	77	3	0	1	0	0	0	0	0	0

年齢(Race) : 20 ~ 80
性別(Gender) : 1 男性、2 女性
人種(Race) : 1 ヒスパニック系ではない白人
2 メキシコ系アメリカ人
3 その他のヒスパニック
4 ヒスパニック系ではない黒人
6 ヒスパニック系ではないアジア人
7 その他
収入(Income): 1 ~15 大きいほど収入大
77、99不明

教育(Education): 1 ~ 5 大きいほど高学歴
5が大学卒業以上
7、9が不明
軍歴 (Veteran): 経験者は1
経験者でなければ9
所帯数 (NOH) : 最大 7
各病歴(症状があった場合 1 なければ 0)
HTN (高血圧)、DM (糖尿病)、IHD (心血管疾患)、CKD(慢性腎臓病)、COPD(慢性閉塞性肺疾患)、CA (ガン)

付録：データセット生成の概要

- 本カップでは[*]を参考にデータを生成しています。[*] では：
 - まず、 NHANES2017-2018から各人（各レコード）が、年齢、人種、収入、教育レベル、軍歴、各種病歴からなるデータセットを抽出
 - 続けて、ガウスコピュラに基づきコロナ重篤化か否かを見積り、その擬似データを上記レコードに付与。つまり、各人（各レコード）が、年齢、人種、収入、教育レベル、軍歴、各種病歴、コロナ重篤化 となる一部合成データとなるデータセット D0 を生成
- 本カップで用いるデータセット
 - 加工前データ D：D0の内、コロナに重篤化した個人をランダムに300名選び加工前データ D とします。
 - 攻撃用データ R：D0のうち、コロナに重篤化しなかった個人をランダムに25名選び、さらにDから25名をランダムに選び合算したデータ R

本カップにおいては上記操作、つまりD0を加工前データ毎に作っています。また、コロナに重篤化する個人の選び方は、原論文よりも高い確率としています。

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

匿名化フェーズ詳細:top2.py

列col でしきい値 theta より大きい行を出力する。列は 1_5 の様に複数与えても良い。

```
for i in range(len(cols)):
    ex &= (df.loc[:, int(cols[i])] < int(thetas[i]))
```

コマンド: python top2.py 加工前:csv 加工後:csv col theta

匿名化フェーズ詳細: bottom2.py

列col でしきい値 theta より小さい行を出力する。列は 1_5 の様にベクトルで与えても良い。

```
for i in range(len(cols)):
```

```
    ex &= (df.loc[:, int(cols[i])] > int(thetas[i]))
```

コマンド: `python bottom2.py 加工前:csv 加工後:csv col theta`

匿名化フェーズ詳細: `kanony.py`

列columnsを準識別子とみなしてk-匿名化する。削除する行を出力

```
def kanony(df, qi=[1, 2], k=1):  
    return df.groupby(qi).filter(lambda x: x[0].count() >= k)
```

コマンド: `python kanony.py 加工前.csv 加工後.csv k col`

匿名化フェーズ詳細: exclude.py

入力input.csvから排除行番号exclude を除いて出力する。

```
exclude_rows = list([int(i) for i in sys.argv[3].split("_")])  
df = df.drop(index=df.index[exclude_rows])
```

コマンド exclude.py 元ファイル名 4 結果ファイル名

匿名化フェーズ詳細: shuffle.py

行と列をランダムサンプリングする。

```
df2 = df.sample(frac=1, random_state=int(sys.argv[3]))  
df2.to_csv(sys.argv[2], header=False, index=False)
```

コマンド shuffle.py 元ファイル名 4 結果ファイル名

匿名化フェーズ詳細: rr.py

列colにおいて、確率 $1 - \text{prob}$ で他の値と置き換える。ただし、
random は乱択アルゴリズムで利用する乱数とする。

```
def rr(x, q):  
    uniq = x.value_counts().index.values  
    y = [i if random.random() < q else random.choice(uniq) for i in x]  
    return(y)  
  
def rrdf(df, q, target):  
    df2 = df.copy()  
    for i in target:  
        df2.iloc[:, i] = rr(df.iloc[:, i], q)  
    return df2
```

コマンド: python rr.py 加工前:csv 加工後:csv prob col random

匿名化フェーズ詳細: dp.py

差分プライバシーに基づいて行colsに ϵ のラプラスノイズを付加する。

```
np.random.seed(int(sys.argv[5]))
df = pd.read_csv(sys.argv[1], header=None)
cols = [int(c) for c in sys.argv[3].split('_')]
epss = [float(e) for e in sys.argv[4].split('_')]
out = sys.argv[2] if len(sys.argv) == 6 else sys.stdout
for i in range(len(cols)):
    df.iloc[:,cols[i]] = lap(df.iloc[:,cols[i]], epss[i])
df.to_csv(out, index = None, header = None)
```

コマンド: `pythondp.py 加工前:csv 加工後:csv col epsilon`

加エプログラム募集 (Python3)

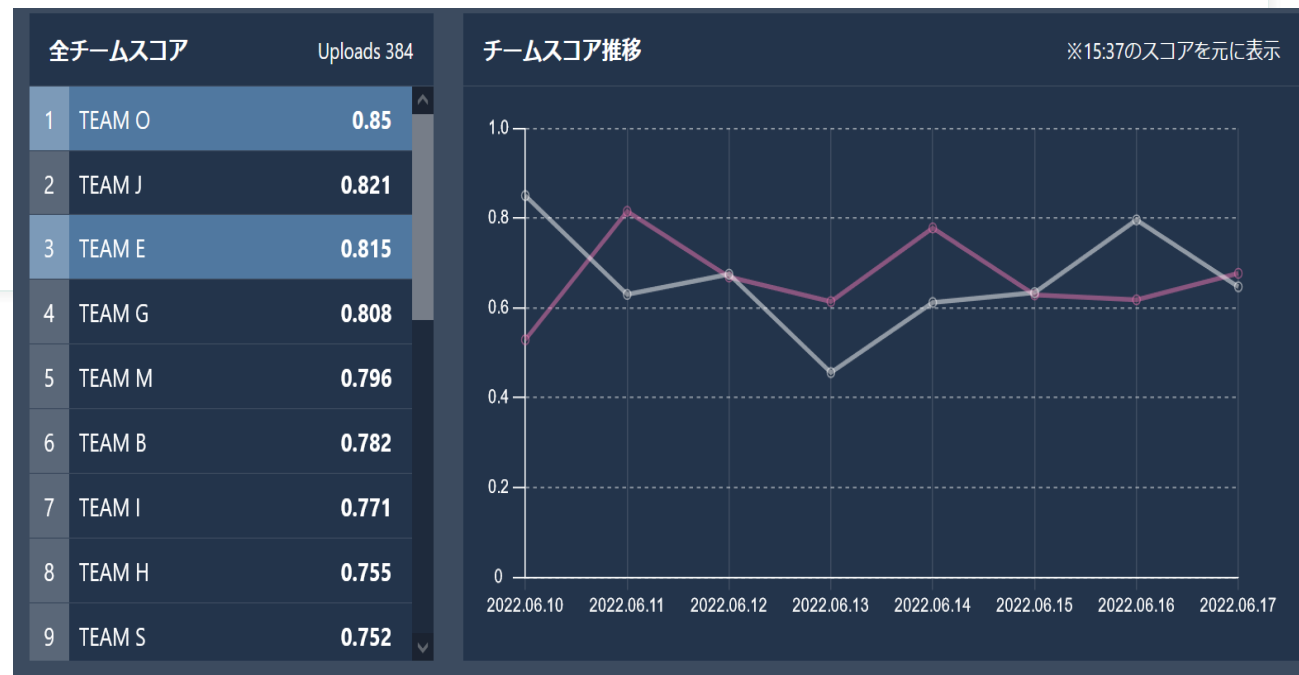
[お問合せフォーム](#)よりお気軽にご提案ください！

匿名化フェーズ詳細 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	2022.06.16 up

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

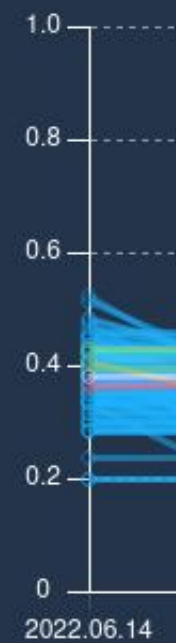
[19]

匿名化ファイルランク

Uploads 99

1	O	dmainO_2_3_2.csv	0.378 [91]
2	G	dmainG_2_2_0.csv	0.367 [91]
3	D	dmainD_2_1_0.csv	0.363 [91]
4	B	dmainB_2_3_0.csv	0.356 [91]
5	E	dmainE_2_2_4.csv	0.356 [91]
6	L	dmainL_2_1_2.csv	0.343 [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



攻撃フェーズ

- 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

攻撃フェーズ詳細 提出物

- 攻撃者は、他の参加者が加工した加工後データD'毎 に下記のような知人リスト R を得ます。
- Rの半分はD' の加工前データDに含まれており、残りの半分はDに含まれていません。

提出ファイル

AGE	GENDER	RACE	INCOME	EDUCATION	VETERAN	NOH	HTN	DM	IHD	CKD	COPD	CA
66	2	4	3	2	0	1	1	0	0	0	0	0
80	2	4	4	3	0	2	1	0	0	0	0	0
76	1	4	7	4	1	1	0	0	0	0	0	0
50	1	6	15	5	0	3	1	0	0	0	0	0
77	2	3	77	3	0	1	0	0	0	0	0	0

知人リストR

1
0
1
1
1
1

参加者（攻撃者）は、Rの各行について、Dに含まれていると判断すれば1、含まれていないと判断すれば0となるファイルを提出してください

総合評価

- **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

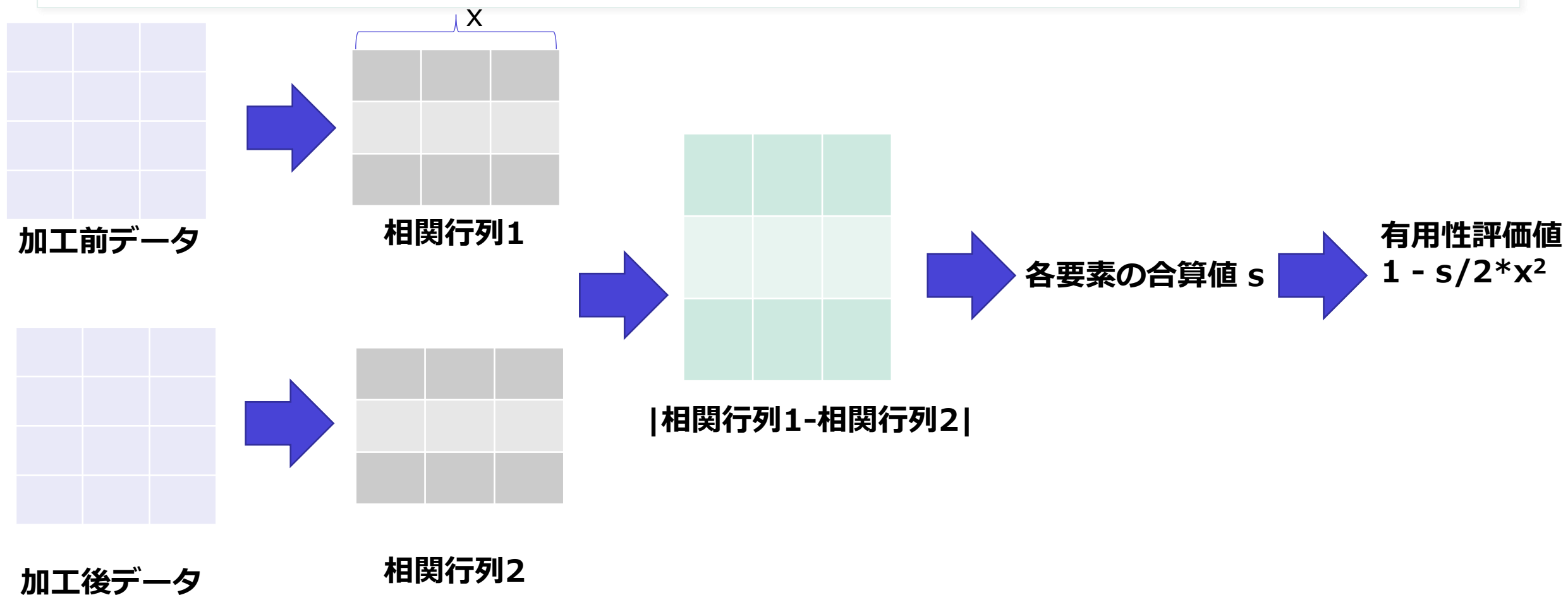
- **総合評価**

- 予備戦 1 対 本戦 9 の比重で評価

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)
- ただし、加工データには重篤化患者しか含まれないため、重篤化患者ではない個人のデータがプログラムに含まれている。

有用性評価：相関行列の差

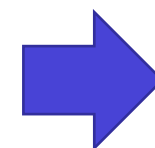


有用性評価：集計数の差

		加工前データ	加工後データ	差分
年齢	20代	23	31	8
	30代	10	11	1
	...	21	23	10
	80以上	12	32	20
	不明 (99)	12	31	19
人種	1 ヒスパニック系ではない白人	32	23	9
	...			
	4ヒスパニック系ではない黒人	31	23	8
	6 ヒスパニック系ではないアジア人	31	33	2
	7 その他	22	1	21
	99 不明	12	3	9
...				
CA	0	44	22	22
	1	11	13	2
	99	1	2	2727272727

合計値A

合計値B



有用性評価値 $1 - B/A$

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。