Fairness, Privacy and Experiment Design

P.S Sommerfelt, R. Syed, R. Akbari, T. Seierstad

December 6, 2021

For reproducibility, all of the material (code, figures, data, etc.) used in this project is available at GitHub .

Contents

1 Introduction **1**

2 Privacy analysis **2**
2.1 Methodology 3
2.2 Utility function 5
2.3 Implementation 6
2.4 Results 7

3 Fair Policies **9**
3.1 Methodology 9
3.2 Implementation 9
3.3 Results 10

4 Experimental Design **13**
4.1 Methodology 13
4.2 Our Policy 14
4.3 Implementation 14
4.4 Results 14

5 Conclusion **16**

6 References **17**

1 Introduction

In this project we will working on the social and scientific aspects of machine learning. We will specially focus on privacy analysis, fairness policies and experiment design. We will discuss various methods for how to make a data set private, i.e. how to protect each individuals sensitive information if the data were public. Then we will discuss fairness in a population and relate to the ethical aspects. We will particularly work with positive Covid-19 individuals who needs treatments in order to recover. Ethical dilemmas often arises

when one has a finite quantity of something and one wants to distribute this quantity according to some rule that can be justified in some sense. This rule or function can be called a utility function. We will assume a finite number of treatments, and our fairness measure [calibration] will take into account the age factor and the severity of the symptoms for distributing treatments. Finally, we will design an experiment for finding an improved policy for distributing treatments.

2 Privacy analysis

An important task in Data Science is privacy analysis. A good data scientist have to protect sensitive data. In this section we discuss the privacy concerns of our data. By having the whole data set public will lead to leaking sensitive information about each individuals, e.g. information about their health, income, age etc. The very same could happen if the whole analysis of data set were online. In order to prevent this, we will use ϵ-differential privacy mechanisms such as randomized response on the categorical data and Laplace mechanism on the continues data. We use these mechanism on the **treatment_data** and **outcome_data**. When we then make the data public, we are guaranteed that none of a individuals private information gets leaked.

When considering how to protect the data in the training set, we will not discuss how to keep the actual data safe, which would require some kind of protected database that can only be accessed by authorized people. We will, on the other hand, focus on how we can protect the data when it is used in the policy and its model. We assess the privacy sensitive features in this data set, to be gender, age, genes, comorbidities and death. What we want to protect is the anonymity of each individual in the training set, even if an adversary has arbitrary side-information, while still retaining almost the same utility [1].

In the imaginary data set there are a lot of individuals, none of whom has exactly the same genes, which is to say that any of the sets of genes in the training set can be used to uniquely identify one of the individuals. However, if you look at each gene in isolation, roughly half of the population has it, and half of it does not have it. What this means is that to preserve anonymity, some privacy measure must be done with the genomic data.

With regards to the other sensitive binary data, gender, comorbidities

and death, several people share the same features, generally because most people in this data set have no comorbidities. However, there still are some sets of features that uniquely identify some of the individuals in the training set, so this part of the data will also need to be anonymized by some stochastic computation.

The ages in the data set are given as floating point numbers, which means that an adversary could theoretically calculate date of birth for each individual, given that they have some information about when the data was collected. There are several ways to anonymize this data, one could put ages within brackets, i.e 0 to 10 years old, 10 to 20 years old and so on, which would anonymize the age data well, but lose a lot of information. Another way to anonymize this data is by using a stochastic function to add some noise to the age data, which is what we have chosen to do, because it gives some privacy without losing too much information. Wang Wu and Hu have done research that suggests that the randomized response outperforms adding Laplace noise when it comes to preservation of utility in a local privacy model[2]. On the other hand, the randomized response mechanism does not work well on floating point number features.

To guarantee that the use of the model does not leak this private information we add Laplace noise to both income and age. On the binary gestures, i.e. the genes, gender and comorbidities, we use the randomized response. The randomized response mechanism can be seen as flipping a coin for each individual feature. If it is heads, return the true value, otherwise flip a new coin; if the new coin is heads, return 0, if it is tails, return 1. The issue with this mechanism is of course that some of the data is 'discarded' and replaced by a random variable, but it is necessary so that an adversary with a lot of side-information still cannot learn any new information from the released policy. We have chosen to use a biased coin, so we don't lose as much of the information.

2.1 Methodology

We will apply the Laplace mechanism in a local private fashion, in other words

$$y_i = x_i + \omega \tag{1}$$

where ω is Laplace distributed with parameter $\lambda = \frac{L}{\epsilon}$. This formula describes how increasing the λ (decreasing ϵ) will lead to increase of the variance of the Laplace distribution. By doing this, we will obtain more privacy, but it also means that we are adding more noise to the data. If we choose a

too small value for ϵ, the data will be close to useless. The process of adding Laplace noise to the data is one of the themes in differential privacy. Looking more closely at the formula one could also understand the meaning of the term ϵ- Differential privacy. The smaller ϵ is, the more information you lose every time somebody is investigating your data. By investigating we mean applying a function to your data in order to extract sensitive information about the the individuals. Larger epsilon means adding less noise. This relation exhibits a trade-off relationship. We want privacy but we also want to have a data set that is useful. We will look closer at this trade-off when presenting the results later.

The λ is also dependent on the the sensitivity of the investigation. The sensitivity is defined as

$$L(u) = \sup_{xNx'} |f(.., x) - f(..., x')| < K|x - x'| \tag{2}$$

One has to define what a neighbor is in the data set(|x-x'|). K is the Lipschitz constant of function f that must be estimated by taking the derivative of the function, setting it equal to zero in order to find the max of the function. Instead of calculating the sensitivity of the main function of interest, the expected utility, we fix it to 1. Then we vary epsilon to get a trade-off between privacy and accuracy.

As mentioned, we use a Randomized response on some of the discrete variables. We have choose to randomize all of the comorbidities, i.e. Asthma, Obesity, Smoking, Diabetes, Heart disease, Hypertension. Further, the Gender variable and the Death variable in the **treatment_data**, and on Death in **outcome_data**. The combination adding noise to all of these variables could however lead to missing too much information. On the contrary, if an adversary want to learn something about the individuals in the dataset she could learn a lot from knowing only the Age and the Gender. Therefore, we wish to add a sufficient amount of noise. For example, one person is much more difficult to find with no information of the gender,death and/or a specific comorbidity. If these variables were public one could just compare these variables with some databases about about death in their specific city to find the specific individual. Therefore, we are randomizing these variables.

The randomized response can be explained by a simple example. In the beginning you flip a coin, in our case a biased coin. If the coin comes up heads, respond to a yes/no question truthfully. If it comes up tails, flip a new coin (non biased). If this second coin turns up heads answer yes and if tails answer no. The first coin we set to have probability (θ) for turning

up heads and $(1-\theta)$ for tails. This means that we have a probability of $\theta + 1/2 \cdot (1-\theta)$ that the answer will be the true answer. We define:

$$\mathbb{E}[p] = (1-\theta)\frac{1}{2} + q\theta \tag{3}$$

where p is the observed rate of positive responses in a sample and q is the true positive response in the population. We can rearrange this equation and find a expression for q:

$$q = \frac{\mathbb{E}[p]}{\theta} - \frac{(1-\theta)}{2 \cdot \theta}$$

The problem with this approach is that we are throwing away some of the data when estimating q. By repeating this procedure at rate ϵ we will obtain that the error bounds would scales as $O(1/\sqrt{\epsilon n})$ for n data points. This means that when n is small, we have a lot of uncertainty on what the randomized observations will be. Any estimate we make using this data set will have much less correlation between the data and the response variable. The randomized response mechanism can be shown to be a $\epsilon-$DP [1] where ϵ is

$$\epsilon = \ln\left(\frac{1-p}{p}\right) \tag{4}$$

Smaller ϵ will as always lead to more privacy. The larger the epsilon the more similar the randomized data will be true data, resulting in no privacy when $\epsilon \to \infty$.

2.2 Utility function

In the following, we present our proposal for the utility function. The symptoms after treatment is associated with a cost. This cost varies by the severity of the symptoms. Our utility function is a sum of the score each individuals gets based on number of symptoms. By summing up all the constants for the symptoms after treatment gives us the score of each person. The utility also have some ethical aspects. We assume that there is a finite number of treatment doses and therefore some individuals have to be prioritized over other individuals. The utility function is constructed in such a way that it prioritizes old individuals over middle aged individuals, who again is prioritized over young individuals. The way we do that is is to give reduced weights to old individuals compared to younger individuals. By penalizing the chance for older individuals to get symptoms less, compared to younger age groups,the policy will decide with a lower threshold to give older individuals treatment over younger individuals since the goal is to maximize the expected utility. In the data set their is not a higher probability for getting

symptoms after getting a treatment across the age groups.

In addition we also have a cost associated to the the treatments. Of course it cost nothing to give the individual no treatment. For treatment 1 and treatment 2 we added a uniform random cost between 0 and 1. It makes sense to not treat somebody with no symptoms. We, therefore, make an assumption in our utility function where we hard-code this assumption by setting the outcome symptoms equal to the symptoms before no treatment given. To these individuals we give no treatment which has a zero cost. A formal definition of our score function is

$$W = \begin{cases} constant \cdot [0, 0, 0.1, 0.1, 0.1, 0.5, 0.2, 0.5, 1.0, 100] & 0 \leq Age \leq 30 \\ costant \cdot [0, 0, 0.2, 0.2, 0.2, 0.7, 0.4, 0.7, 2.0, 90] & 30 \leq Age \leq 60 \\ constant \cdot [0, 0, 0.5, 0.5, 0.5, 1.2, 0.5, 1.0, 10.0, 70] & 60 \leq Age \end{cases}$$

$$R = W \cdot y$$

R is a vector where each component is scaled by the its respective weight and y is the symptoms of the individuals after a specific action is applied. The score is then given by

$$S(a, y) = \sum_i R_{i,a} - cost_{i,a}$$

where a is specific action_type. The equation above represents the total score for a specific symptom. For a fixed action, the utility becomes:

$$U(a = a, y) = \sum_y S(a, y)$$

2.3 Implementation

We have implemented the randomized response in the same manner as discussed above. We randomly chose n *flips* with a parameter θ. Then we randomly select noise variables with n length. Then we flip those data points where the flip were false. The Laplace mechanism is implemented by the equation 1, e.g. adding a noise constant from Laplace mechanism to the actual data. In order to check that our randomized technique has worked properly, we have used a bootstrap method to display the distribution of the data with and without the added noise. Furthermore, we have tuned the θ parameter in order to find a trade off between accuracy between the noisy data and original data and the ϵ.

θ	ϵ	loss in utility
0.500	0.000	1140600
0.549	0.197	1031600
0.598	0.397	910200
0.647	0.606	807300
0.696	0.828	693600
0.745	1.072	569400
0.794	1.349	470850
0.843	1.681	365500
0.892	2.111	250900
0.941	2.769	147250
0.990	4.595	26950

Table 1: Showing the estimate for the amount of loss in utility as the privacy guarantee varies. Here we have used the absolute difference between expected utility for the noisy data and the original data. The values of ϵ is computed by 4.

2.4 Results

As expected, we obtained bigger loss for smaller values of the ϵ. This is shown at 1. At $\epsilon = 0$ we have perfect privacy but also accuracy at 0 and hence bigger loss in the expected utility. While for the bigger ϵ values we get less loss values.

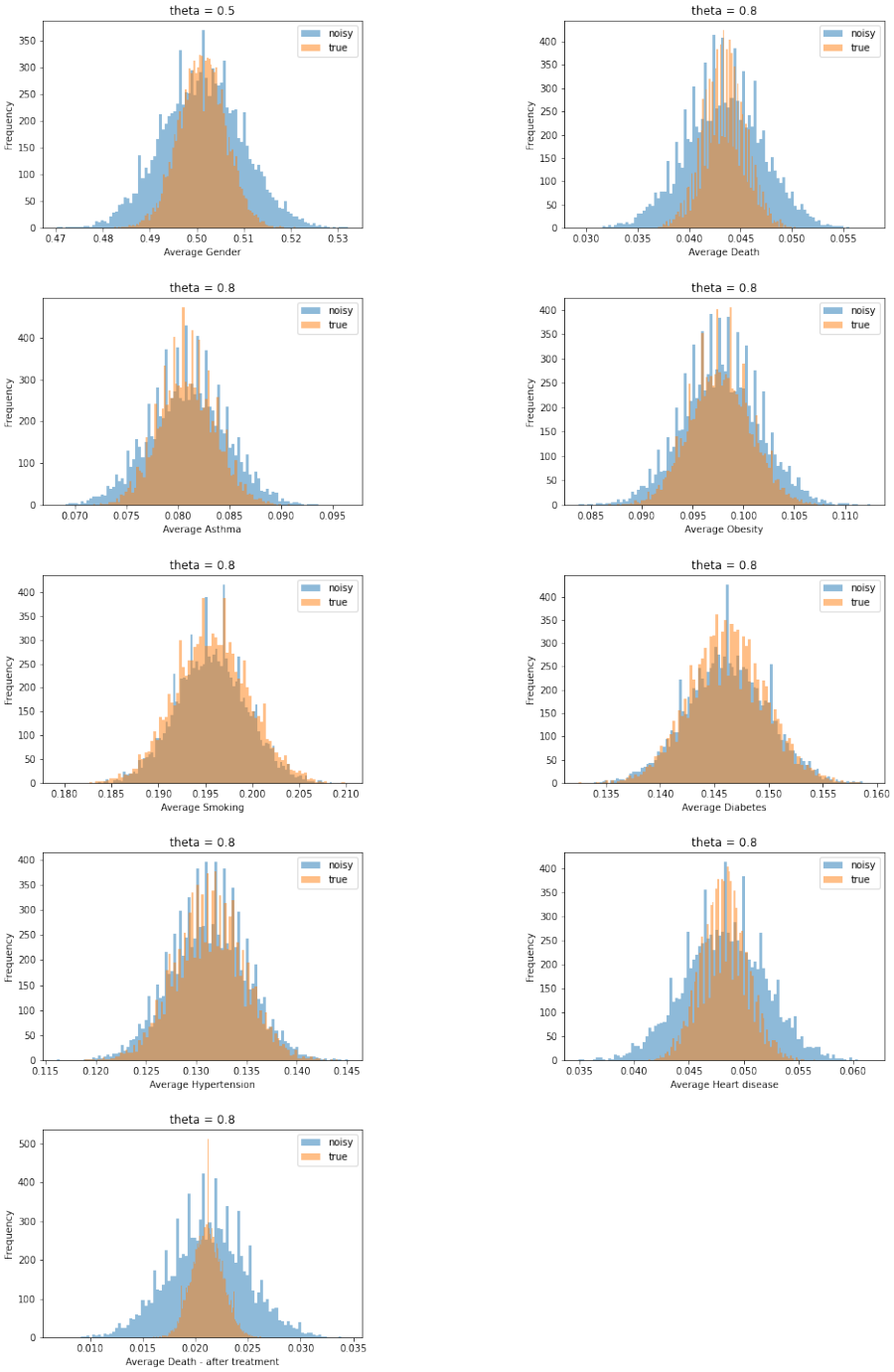

Figure 1: Distribution of the data with and without added noise after randomized response. Here we have chosen $\theta = 0.5$ for Gender and $\theta = 0.8$ for every other symptom.

3 Fair Policies

3.1 Methodology

We have decided to declare the age as the sensitive variable. We dichotomize the age variable into two categories based on mean of the ages. Those samples with ages below the mean of age gets annotated as 0 and the others as 1. Furthermore, we are interested in whether our policy from above is fair. To measure the fairness of the policy π, we have chosen to use calibration

$$\mathbb{P}^{\pi}(y|a,z) = \mathbb{P}^{\pi}(y|a)$$

where a is the treatments given to the patients, z is the sensitive variable and y is the outcome of the actions on the patients. In order to calculate the fairness measure, we use

$$\sum_{i} \left| P_{\theta}^{\pi}(y|a,z) - P_{\theta}^{\pi}(y|a) \right| \tag{5}$$

We want this measure to as low as possible for all outcomes in order to declare it fair. By studying the original features, we can observe that the some biased variables are included. Especially, the income variable. All the other variables explain something about the health of each individuals except income variable which makes it biased.

3.2 Implementation

We are interested in finding how the treatments a effects the symptoms. To visualize this, we have used bootstrap methods to generate 1000 samples of the, then grouped by the sensitive variable and taking the mean of the data for each symptom.

We have chosen multilayer perceptron (MLP) classifier in section. We first train our classifier with action data a and sensitive variables z as input and outcome data y as response. Then we train a new classifier again with action data without sensitive variables as input and outcome data y as response. In order to find the probabilities from equation [1], we have used Sckitlearn's ***predict_proba*** function.

3.3 Results

Symptoms	Fairness measure
Covid-Recovered	6.638951
Covid-Positive	0.685614
No-Taste-Smell	22.057378
Fever	47.492659
Headache	1.669986
Pneumonia	69.627222
Stomach	3.624037
Myocarditis	5.510939
Blood-Clots	28.667796
Death	30.763775

Table 2: Showing the fairness of each symptoms in the historical data.

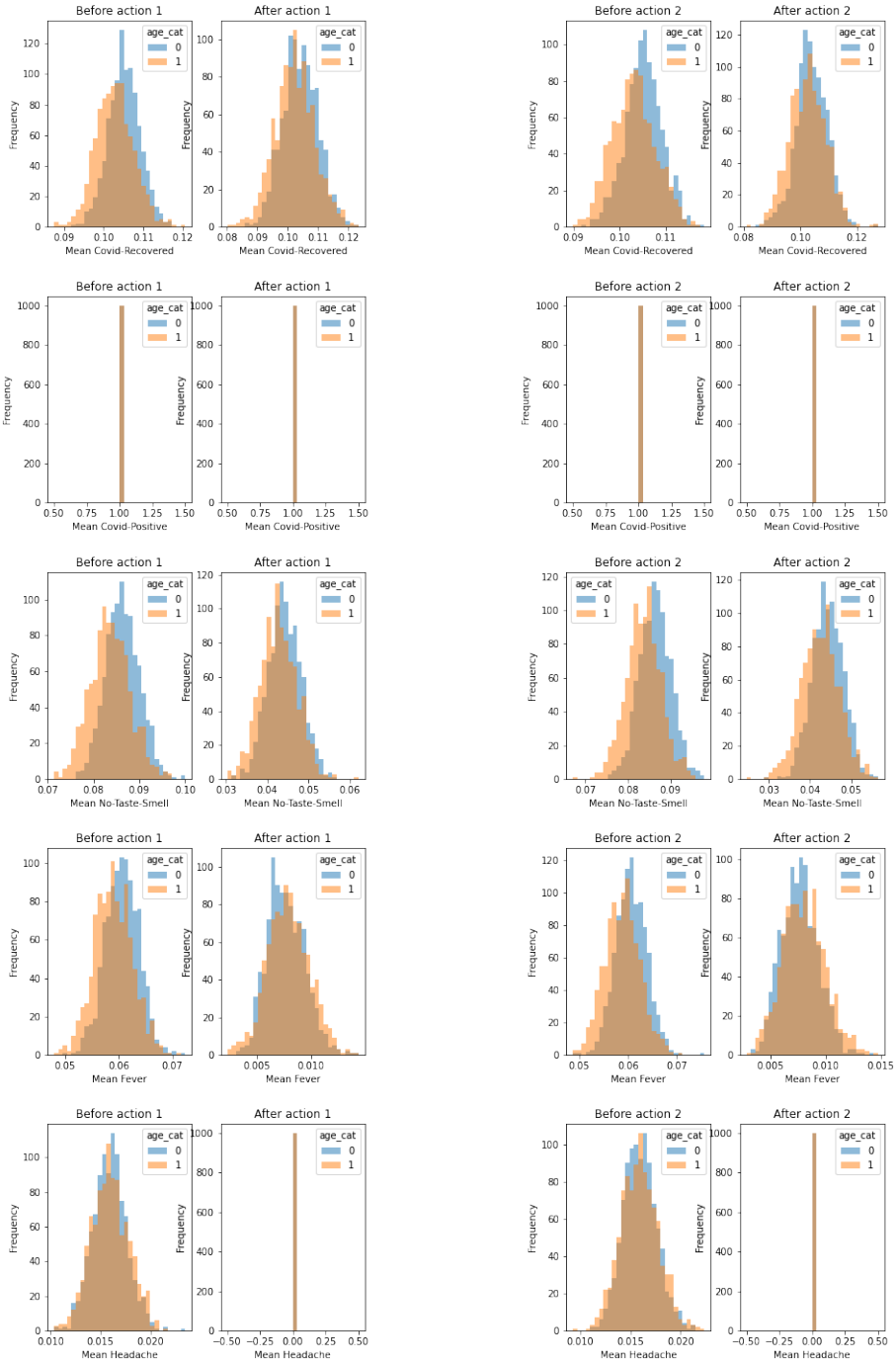

Figure 2: Distribution of hte mean of 1000 bootstrapped sample of the population with respect to Covid-Recovered, Covid-Positive, No-Taste-Smell, Fever and Headache. The plots are showing the each symptoms before and after the actions, e.g. 1 or 2.

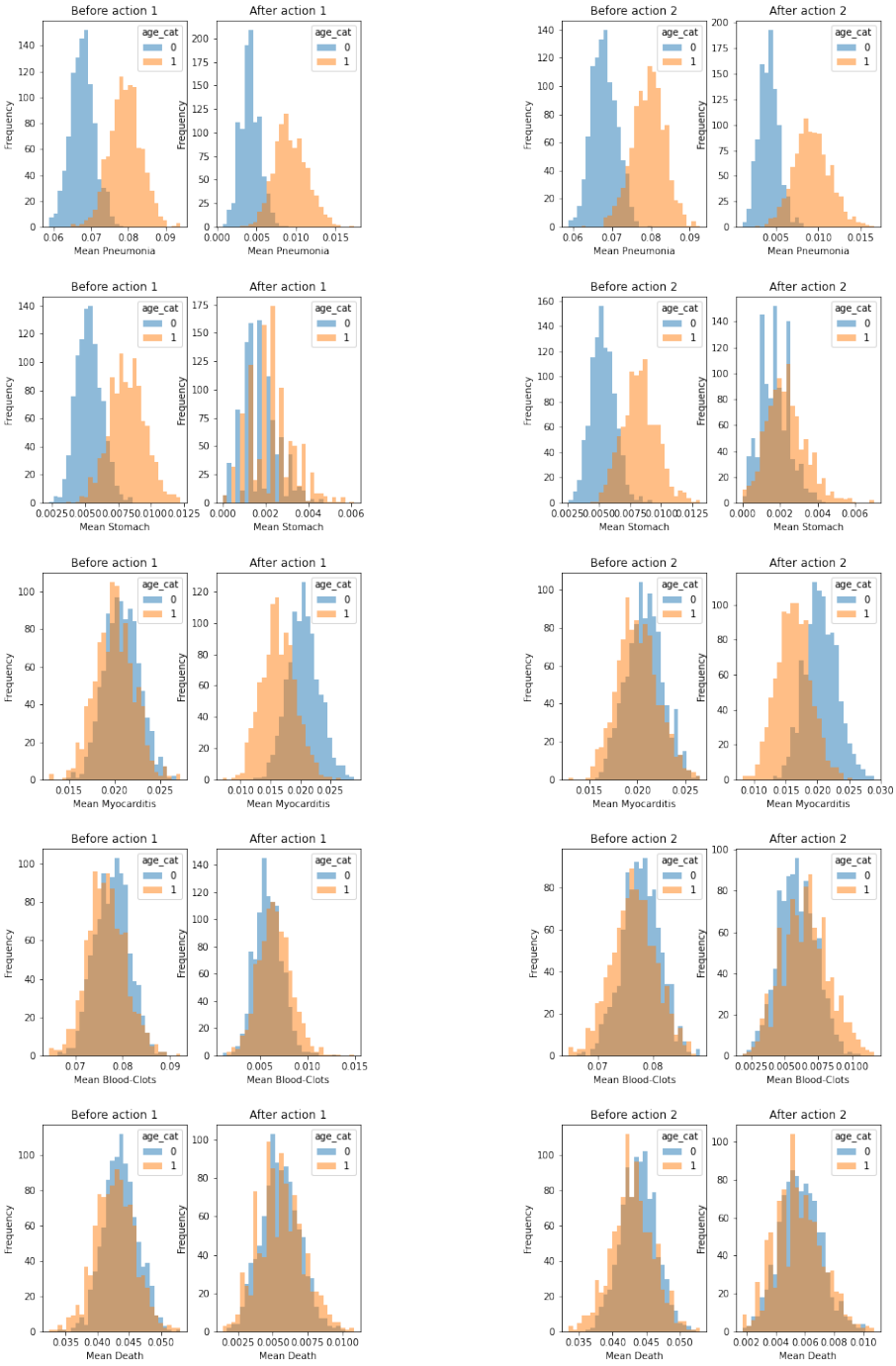

Figure 3: Distribution of mean of 1000 bootstrapped sample of the population with respect to Pneumonia, Stomach, Myocardities, Blood-Clots and Death. The plots are showing the each symptoms before and after the actions, e.g. 1 or 2.

The figures at 2 and 3 shows how the symptoms are distributed for each sensitive variable for each action. From these plots, we can draw some conclusions about the data. For variables like No-Taste-Smell and Pneumonia decrease a bit after both actions are applied. The distribution for the Headache, in both cases of the actions, goes towards zero which implies that the given actions works perfectly for this symptoms. However, the mean for the distributions increase a little bit after the given actions. This could be some sort of side effects of the treatments. The most important variable for our study is Death, and it looks the actions works good against this symptom. But it is not perfect since the mean of the distributions are going towards zero after the applied actions.

The table [1] above shows the initial fairness of the symptoms using the random policy. As mentioned above, we want this fairness values to be as low as possible. We can for example see for symptoms like *Headache* that the fairness value is very small which means that the patients with this symptoms get very fair treatment regardless of age category.

4 Experimental Design

In this section we will focus on the experimental design. Using the utility function that we proposed earlier, we will find the expected utility of the historical policy and present an improved policy.

4.1 Methodology

We define the expected utility of an action to be

$$\mathbb{E}(U) = \sum_{y} U(x, a, y) p(y|a, x) \tag{6}$$

where U is the utility function, a is the given treatment, the x's are the features of patients before receiving any treatment, the y's are the symptoms after receiving treatment, and $p(y|a, x)$ is the probability of having a symptom after receiving treatment a for a patient with features x.

Since we do not know the outcome of our policy yet we use $p(y|a, x)$, estimated on the historical data using a machine learning algorithm, to simulate a new data set for the outcome. That is, for the patients receiving one treatment we estimate the probabilities of having symptoms after finished treatment and use these probabilities to simulate the outcome. This is done

so that we are able to estimate the expected utility of our proposed policy as our utility function depends heavily on the outcome data.

4.2 Our Policy

Our proposed policy is to find the optimal action for each person in order to maximize the expected utility. In our case, we want to maximize $\mathbb{E}(U)$ with respect to the actions. The action chosen by our policy becomes

$$a^* = argmax_a\{\mathbb{E}[U(a,y)]\} = argmax_a\Big\{\sum_y p(y|a,x)\cdot U(x,a,y)\Big\}$$

4.3 Implementation

We start by dividing the data into two groups based on Treatment1 and Treatment2. Then we calculate $P(Y|a = Treatment1, X)$ and $P(Y|a = Treatment2, X)$. Furthermore, we calculate the utility for each data based on each action types. The expected utility is obtain by using the 6. Depending on the treatment that gives the best expected utility, we choose to give actions to the specific individual. Based on this optimal actions, we pick out the individuals (improved data) that was given these actions.

By taking the improved and the previous probabilities for each action types, we simulate new outcome data. These outcome data are made under the improved policy. Furthermore, we continue the same process as mentioned above for finding the expected utility.

4.4 Results

As listed in table 3, we can see that our improved policy has slightly better average expected utility. By average expected utility we mean expected utility divided by number of individuals within that action category. In other words, this table shows that improved policy is better than the historical policy. As mentioned above, we had defined the utility function with a finite number of treatments. We have to take into account the ethical aspects in order to distribute the treatments among the individuals fairly. We defined the severity of the symptoms and the age categories as the main parameters for choosing whether a person got a treatment or not. Figure 4 illustrate this idea. We can see that number of people that got treatments has decreased evenly throughout the distributions. Interestingly, the amount of Action2 has decreased a lot more compared to Action1. However, this confirms that our approach for utility works. We do not want to treat individuals with no

symptoms, as the treatments are not showing to have any effect on treating Covid-19, only the symptoms that comes with the disease. And since we are giving a small cost to the treatments our policy chooses, as expected, not to treat patients with no symptoms.

Expected utility	Action 1	Action 2
Historical	-1.5033	-2.0902
Improved policy	-1.4982	-1.6635

Table 3: The expected utility value for each individual for both historical data and the simulated data with improved policy. They are separated based on the action types.

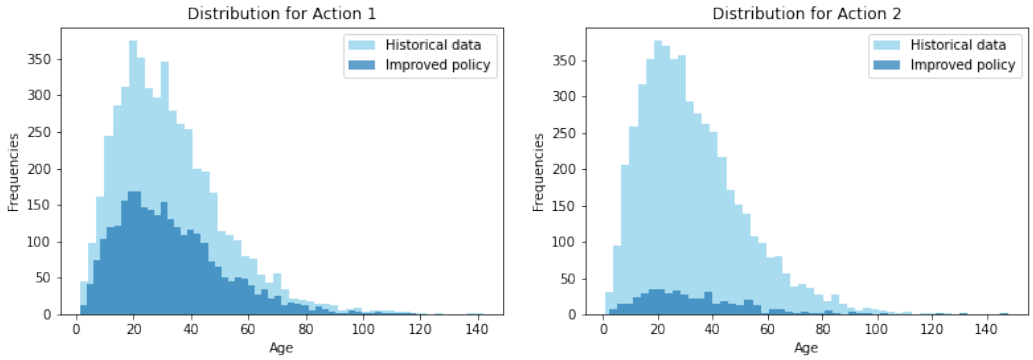

Figure 4: Showing distribution of Age variable for Action 1 and Action2.

We also take into account the error bounds for expected utility. We have solved this issue with confidence intervals. We assume that

$$\frac{\mathbb{E}(U) - \overline{u}}{\frac{s}{\sqrt{n}}} \sim t_{n-1} \tag{7}$$

where s is the sample standard deviation and $\overline{u}$ is the sample mean of the utility.

The sample variance is defined as

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{n} (u_i - \overline{u})^2$$

and the standard deviation is $s = \sqrt{s^2}$

This results in this confidence interval with confidence level α

$$\mathbb{E}(U) \pm t_{1-\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$$

On the historical policy, we construct a 95% confidence interval for each the expected utility of each individual receiving treatments:
Treatment 1:

$$(-1.2790708252994114, -1.7272383432280067)$$

Treatment 2:

$$(-1.5390055863581198, -2.6414226158170946)$$

For the improved policy we got:
Treatment 1:

$$(-1.2425517737387357, -1.5219074725668047)$$

Treatment 2:

$$(0.4553587127584242, -2.9555947121045985)$$

From this we see that our proposed policy has made significant improvements to the expected utility for the people obtaining Treatment 2. However, since fewer people was given treatment under our policy, the variances of expected utility are also higher which gives us less confidence in our estimates.

5 Conclusion

We started with implementing ϵ-DP mechanism, e.g. Randomized response for categorical variables and Laplace mechanism for continuous variable. We chose the to run the mechanism on these variables; Gender, Death,Asthma, Obesity, Smoking, Diabetes, Heart disease, Hypertension, Age in the **treatment_data**, and the death variable in the **outcome_data**. We tuned the ϵ parameter by running the randomized response with different θ values. We obtained that for smaller ϵ we got more loss and vice versa. Our utility was defined to be a trade off between prioritize the older individuals in the population and the cost of treatment,since we assumed that we had a finite amount of treatment doses. The policy we choose was to find the action that maximized expected utility.

In order to do fairness measure, we first carried out a simple data analysis where we analyzed how the different actions affected the different symptoms. We saw for example both actions minimized the Headache variable perfectly to zero while it decreased Death but could not make it to zero. For the fairness measure we used calibration. Estimating the fairness of outcome for each outcome symptom, gave us some clues into how fair the historical data was. Our sensitive variable was age. The empirical distribution looked pretty similar across all symptoms. We found out that the most fair treatment were given to Headache and Stomach, which could be seen from the plots (distributions drawn towards zero).

In the last section of the project, we designed an experiment for finding estimate of the utility for historical data and then found an improved policy based on the previous knowledge. As table 3 shows, the improved policy had better expected utility. Since we had defined a cost to each treatment our policy does not treat people with no symptoms. We had decided to prioritize individuals based on their severity of symptoms and age. That was done to take into account the ethical aspects. The result of this were plotted in the 4. This showed how much of the distributions were reduced from the historical data. To take care of error bounds, we introduced and computed confidence interval for expected utility and concluded with our estimates were good estimates.

6 References

[1] Chapter 3.4 in Machine learning in science and society,
(https://github.com/olethrosdc/ml-society-science/blob/master/notes.pdf).

[2] Using Randomized Response for Differential Privacy Preserving Data Collection, http://csce.uark.edu/ xintaowu/publ/DPL-2014-003.pdf