IN-STK 5000 Project 1

Meirav, Anders, Thomas

October 24, 2021

1 Task 1

1.1 a) Predicting the effect of features on symptoms

Preprocessing: First, we construct a binary feature for two age groups (< 55 vs ≥ 55). Second, while many combinations of features are possible, in practice we expect to see only few frequent combinations. As a preprocessing step, we use the A-Priori algorithm (with minimal support set to 0.05) for frequent item set mining[1]. We will use the output of this algorithm to construct a new binary feature for each frequent combination of features. We limit the combinations to size 2, but this could easily be extended. For labels (symptoms) with low frequency in the data (set to be lower than 0.05 of the samples) we perform balancing (under sampling the negative class - using only half of the samples and oversampling the positive class - sampling 0.05 of the original number of samples).

Feature selection pipeline: We test Lasso regression with age and comorbidities as features, and symptom as the label (we look at each symptom separately). To select the best hyperparameters we perform k-fold cross validation ($k = 5$). The selected hyperparameter (evaluated by f1 score of the model). Then the model is trained on the entire dataset and used to extract meaningful features (non-zero weights of the model). After finding the interesting features, we use the weight given to each feature as a measure of the effect.

Results are presented in figure 1, there we can see that only combinations of features are given a positive weight. Note that we only have results for two symptoms, since the models for the other symptoms were not able to produce a reasonable outcome due to sparse symptoms.

To estimate the robustness of this pipeline, we first use simulated data.

Data generation: We simulate data according to the features' distribution in the given dataset. For binary features we use Bernoulli distribution, using the mean of the feature in the data as the distribution parameter. For age, we use Gamma distribution with $a = 1.99$ and std based on the original data. The simulated data is constructed under the assumption that features are not correlated, which could be extended in the future.

The target bit (simulating a symptom) is simulated using different levels of influence by an arbitrary selected feature, in our case Asthma. The meaning of effect 0.7 is that for every individual with Asthma, the target bit was set using a Bernoulli distribution with parameter 0.7. The baseline distribution was based on the mean of the feature in the original data. This was simulated for the different symptoms in the data. For each combination of symptom and effect

[1] https://github.com/chonyy/apriori_python

size, we perform the pipeline describe above with 5 repetitions. In table 1 are the results showing how many times on average was the right feature (Asthma) found to be significant by the model. It seems that there are differences between the different symptoms, and it is easier for the model to find the right feature when the effect is greater. In table 2 we can see the average number of non-zero weight features in the trained model. In this case, since there are no added effects, the result of 1 would be optimal. We get this result for effect of 0.9. In other cases, the model identified additional features as significant, which are in fact only noise.

Table 1: Average number of times the right feature was selected as significant by the model, for different effect size and different symptoms

Effect	0.1	0.3	0.5	0.7	0.9
Covid-Recovered	0	0.6	1	1	1
Covid-Positive	0	0	1	1	1
No-Taste/Smell	1	1	1	1	1
Fever	0	0	1	1	1
Headache	0	1	1	1	1
Pneumonia	1	1	1	1	1
Stomach	1	1	1	1	1
Myocarditis	1	1	1	1	1
Blood-Clots	1	1	1	1	1
Death	1	1	1	1	1

Table 2: Average number of features selected as significant by the model, for different effect size and different symptoms.

Effect	0.1	0.3	0.5	0.7	0.9
Covid-Recovered	0	2.8	3.8	1.6	1
Covid-Positive	0	0	7	4.6	1
No-Taste/Smell	6.6	3.2	1	1.2	1
Fever	0	0	6.6	1.2	1
Headache	0	2.2	1	1	1
Pneumonia	4.4	3.2	1	1.2	1
Stomach	4	2.4	1	1	1
Myocarditis	2.2	2	1.2	1.8	1
Blood-Clots	4.8	2	1	1	1
Death	5	2	1	1.4	1

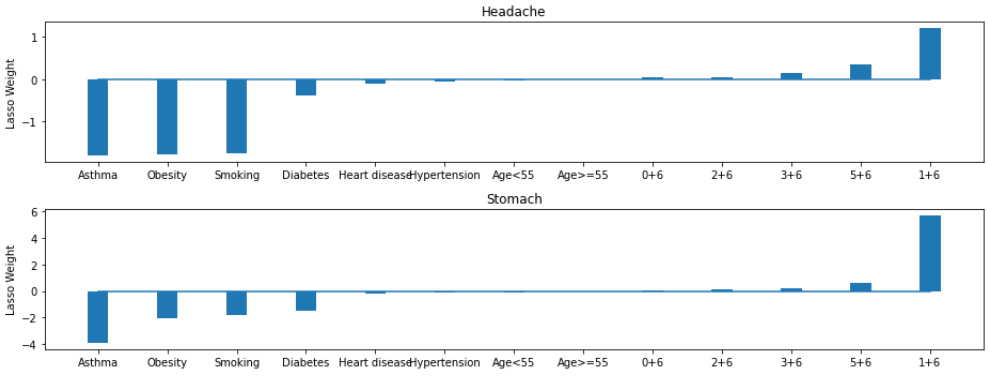

Figure 1: Lasso logistic regression weights for features

1.2 b) Estimating the efficacy of vaccines

We consider a data set D comprising feature vectors $\mathbf{x}_i$ corresponding to individuals $i \in [N]$, where features are numerical or categorical. We use the following notation: $\mathbf{x}_i(cov)$ equals 1 if individual i is Covid-positive, and 0 if individual i is Covid-negative.[2] Similarly, $\mathbf{x}_i(vac_j) = 1$ if individual i is vaccinated with the j-th vaccine with $j \in [3]$.

In this section, we investigate the efficacy of vaccines, i.e. we want to determine to what degree certain vaccinations can prevent an infection with Covid. We define the *risk* of a Covid-infection of group A as the empirical frequency[3]

$$risk_{covid}(A) = \sum_{i \in A} \frac{\mathbb{1}(\mathbf{x}_i(cov) = 1)}{|A|}. \tag{1}$$

We are interested in the efficacy of each of the three vaccines. To this end, let $\mathcal{N}(\neg vac)$ denote the group of individuals that are unvaccinated. Similarly, let $\mathcal{N}(vac_j)$ denote the set of individuals that have taken vaccine vac_j with $j \in [3]$. The Vaccine Efficacy (VE) of a vaccine vac_j is then defined as the proportional reduction in cases among vaccinated individuals, i.e.

$$\text{VE}(vac_j) = \frac{risk_{covid}(\mathcal{N}(\neg vac)) - risk_{covid}(\mathcal{N}(vac_j))}{risk_{covid}(\mathcal{N}(\neg vac))}. \tag{2}$$

We interpret $\text{VE}(vac_j)$ as the proportionate reduction of infections in the group vaccinated with vaccine vac_j. For instance, if in a population of 100 unvaccinated individuals and 100 vaccinated individuals, we observe 25 and 1 Covid-infection in the unvaccinated and vaccinated group, respectively, the vaccine efficacy is given by $\text{VE} = (25 - 1)/25 = 0.96$, i.e. a reduction of infections by 96%.

Challenges. It may be the case that vaccinations are differently effective across demographic groups. For instance, it may be the case that a vaccine is highly effective when given to male patients, however, ineffective when given to female patients. To get better insight into the

[2]Note that we do not consider the feature 'Covid-recovered', but only 'Covid-positive' as we do not know whether a Covid-recovered individual has been vaccinated before or after their Covid-infection. On the other hand, with the limited information that we have about the data, we assume that individuals that are Covid-positive and vaccinated received the vaccine before their infection with Covid.

[3]We base our definitions of risk and vaccine efficacy on www.cdc.gov/csels/dsepd/ss1978.

actual vaccine efficacy of a certain vaccine we must account for demographic groups and vaccine peculiarities.

In general, since we can only conduct a *non-random* study, another factor that we should consider is the (unknown) decision rule by which vaccinations were distributed among e.g. different age groups. This could, for instance, be done by matching individuals from the unvaccinated and vaccinated groups according to their feature similarity. However, we will refrain from using a complicated matching procedure, as a quick look into the data suggests that vaccinations were handed out uniformly at random.

Adjusting by Age, Gender, etc. In order to identify crucial peculiarities of the vaccines, e.g. that a vaccine is less effective when given to female patients, we stratify the data into different groups and analyse its vaccine efficacy for each group separably. We adjust according to the feature Age (< 55 vs ≥ 55) and Gender (male vs female). For each group $\mathcal{N}(\neg vac)$, $\mathcal{N}(vac_1)$, $\mathcal{N}(vac_2)$ and $\mathcal{N}(vac_3)$, we extract the individuals with e.g. Age < 55 and compare the vaccine efficacy using the empirical frequencies of Covid-infections among the group, i.e. the risk of an infection (cf. equation (1)).

Computing Risk and Vaccine Efficacy. For each vaccine and adjusted group (e.g. Age $<$ 55), we compute the vaccine efficacy $\text{VE}(vac_j)$ according to equation (2). The estimated vaccine efficacy is reported in Figure 2. In addition, we report the risk of an infection, i.e. the empirical frequency given by equation (1), including .95 confidence intervals, which were computed using Hoeffding's inequality. Note that the confidence intervals relate to the number of patients used to compute the empirical frequencies. The more patients were considered when computing the empirical frequencies, the tighter the intervals. In addition, we also include the number of patients in each adjusted group in Figure 2. Ideally, we would report confidence intervals for the vaccine efficacy as well, however, it is not clear (to us) how to compute them as we divide by the empirical frequency $risk_{covid}(\mathcal{N}(\neg vac))$ when computing $\text{VE}(vac_j)$.[4]

Interpretation of Results. We report the vaccine efficacy and risk for all three vaccines in Figure 2. We see that generally Vaccine 1 is the least effective with only a 11.3% reduction in Covid-cases overall. In particular, Vaccine 1 appears to ineffective for ages ≥ 55 as the vaccine efficacy is given by roughly 4.46%.[5] The sex of patients appears to have no effect on the efficacy of the vaccine. Vaccine 2 shows to be more effective than Vaccine 1; in particular showing effectiveness (23.6%) in age group ≥ 55. Vaccine 3 is the most effective vaccine with vaccine efficacy of roughly 30% across all ages and sexes (slightly lower efficacy of 28.7% for age group ≥ 55. Overall we must conclude that all three vaccine are relatively ineffective as the expected reduction in Covid-cases has been seen to be lower than or equal to 30% for all vaccines. In comparison, the Biontech vaccine has been reported to have a vaccine efficacy of roughly 72%.[6]

[4] Question at Christos or whoever is reading this: Is it possible to use Hoeffding's inequality for the vaccine efficacy? Clearly, without dividing by $risk_{covid}(\mathcal{N}(\neg vac))$ and simply taking the difference $risk_{covid}(\mathcal{N}(\neg vac)) - risk_{covid}(\mathcal{N}(vac_j))$, we could still apply Hoeffding's inequality by defining a new RV that is just the difference of the two indicators. However, after briefly thinking about it, we don't see how one could handle the denominator.

[5] Note that even though there are only 2601 vaccinated (and 5290 unvaccinated) individuals in the age group ≥ 55, the results still appear to be significant as the confidence intervals around the risk appear very small.

[6] https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(21)00224-3/fulltext

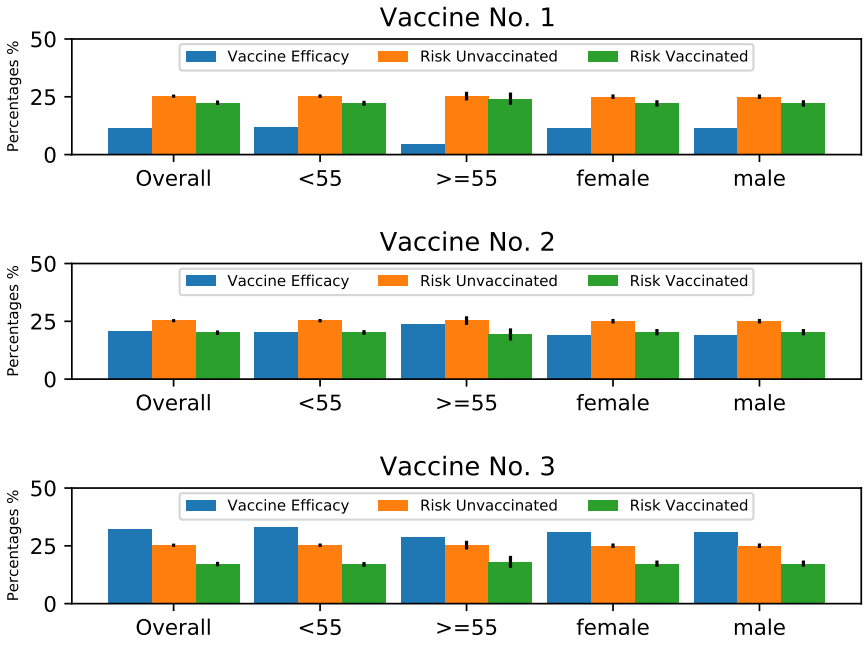

Figure 2: Vaccine Efficacy (see eq. (2)) and Risk (see eq. (1)) adjusted for Age and Gender. Unvaccinated individuals have 'overall' risk 25% $\approx$ 10122/40036, for Age $<$ 55 risk 25% $\approx$ 8865/35074, for Age $\geq$ 55 risk 25% $\approx$ 1339/5290, for female patients risk 25% $\approx$ 4995/19935, and for male patients risk 25% $\approx$ 4995/19935. Exemplary, Vaccine 1 has 'overall' risk of 22% $\approx$ 4431/19758, for Age $<$ 55 risk 22% $\approx$ 3846/17310, for Age $\geq$ 55 risk 24% $\approx$ 629/2601, for female patients risk 22% $\approx$ 2187/9875, and for male patients risk 22% $\approx$ 2187/9875.

1.3 c) Estimating the probability of vaccination side-effects

Method The provided data set includes 3 types of vaccines, where no combinations of them are present. We therefore analyse them separately in the pipeline. Of the 10 symptoms in the data set we regard the following 8 as side-effects of the vaccines: No-Taste/Smell, Fever, Headache, Pneumonia, Stomach, Myocarditis, Blood-Clots and Death.

With respect to Covid-positive and Covid-recovered we removed all individuals that had experienced either of these diagnoses in this section due to the following reasoning. There is no information of the timeline of the different observations on the individuals, but we reason that the side-effects were reported in a certain time-window after vaccines were given. Further, that Covid-positive overlapped with this window and that Covid-recovered had had Covid prior to this time window. Both Covid and vaccines can influence the side-effects and vaccines can influence the prevalence of Covid. Regarding Covid-recovered, there might be long lasting effects of Covid infections that influence the symptomes. This could be evaluated by comparing Covid-diagnosed with and without vaccine, but potential synergistic effects can not be controlled. Thus, the safest control group is individuals without any Covid diagnosis.

We used logistic regression from the statsmodels v0.13.0 packaged to fit models to the data. One

model for each combination of one vaccine and one side effect was fitted. Thus, each model will take in all observations for a specific vaccine (vac_j) together with all unvaccinated in combination with a particular side-effect ($side\text{-}effect_k$). j is the 3 different vaccines and k denotes the 8 side-effects. $vac_j = 1$ and $side\text{-}effect_k = 1$ when the features are present. From the models we predicted the various probabilities of the various side-effects given any acquired vaccine, i.e.

$$Pr(side\text{-}effect_k | vac_j).$$

To compute a confidence intervals of the probabilities we firstly calculated the Wald confidence interval for the linear combination in the logistic model ($x^T\beta$, where x represent presence of the vaccine feature and β the estimated coefficients of the regression) using

The estimated standard error (SE) is calculated from the covariance matrix of the regression coefficients (Σ) by

$$SE(x^T\beta) = \sqrt{x^T \Sigma x}$$

Thereafter, we applied the logit transformation on the linear confidence interval to get the upper and lower bounds of the logistic probability confidence interval, i.e.

$$\frac{e^{x^T\beta - z \cdot SE(x^T\beta)}}{1 + e^{x^T\beta - z \cdot SE(x^T\beta)}}, \frac{e^{x^T\beta + z \cdot SE(x^T\beta)}}{1 + e^{x^T\beta + z \cdot SE(x^T\beta)}}.$$

z is set to 1.96 to achieve a 95% confidence interval.

In situations when there are zero observations of a particular vaccine-symptom combination (as for vaccine 3 and death) the Newton method, which is used to find the maximum likelihood estimates, will not converge. This is due to perfect separation of the categories by a hyperplane. Thus, our pipeline will give out zero probability (with no confidence interval) before trying to fit a logistic regression model in such situations.

$$x^T\beta \pm z \cdot SE(x^T\beta).$$

Interpretation of Results The output from the pipeline with the results of probability estimates of the effect of vaccines on side-effects is given in Table 3. Most probabilities are below 1 percentage and the 3 vaccines show quit similar results. For symptoms of less severe diseases such low probabilities is of less importance. However, there are some exceptions. All vaccines appear to give an 10% probability of fever and 6% for Headache. Furthermore, regarding the serious diseases blood clots (0.51%) and death (0.07%), vaccine 2 show significantly higher probabilities then the two others. For the other serious disease, myocarditis, vaccine 3 showed significantly larger probabilities (0.53%) compared to the other vaccines. In conclusion, vaccine 1 appear to be the safest of the tree vaccines.

Table 3: The estimated probabilities of the effect of the tree vaccines on the various side-effects. Values are given in percentage and 95% confidence intervals is denoted in brackets.

	No-Taste	Fever	Headache	Pneum.
Vac 1	0.09 (0.05, 0.15)	10.34 (9.85, 10.84)	5.77 (5.41, 6.16)	0.12 (0.08, 0.20)
Vac 2	0.08 (0.04, 0.14)	10.28 (9.81, 10.78)	5.81 (5.45, 6.19)	0.18 (0.13, 0.27)
Vac 3	0.04 (0.02, 0.09)	10.61 (10.14, 11.10)	6.01 (5.65, 6.39)	0.08 (0.04, 0.13)

	Stomach	Myoc.	Blood-Clots	Death
Vac 1	0.23 (0.16, 0.32)	0.05 (0.02, 0.10)	0.15 (0.10, 0.23)	0.02 (0.01, 0.06)
Vac 2	0.20 (0.14, 0.28)	0.06 (0.03, 0,11)	0.51 (0.41, 0.63)	0.07 (0.04, 0.13)
Vac 3	0.30 (0.23, 0.40)	0.53 (0.43, 0.65)	0.10 (0.06, 0.16)	0.00 (0.00, 0.00)

2 Task 2

We model the effect of treatments on alleviating symptoms based on an observational study. To this end, we are being provided with data sets *treatment_ features.csv*, *treatment_ action.csv*, and *treatment_ outcomes.csv*. We will refrain from discussing the data in detail as a proper description of the features, treatments, and outcomes is provided in the task description.

Formalising the Problem. Let $\mathcal{X}$ denote the features space, i.e. each row in the data set *treatment_ features.csv* is element in $\mathcal{X}$. Similarly, let $\mathcal{A}$ denote the action space corresponding to data set *treatment_ action.csv* consisting of he treatments prescribed to each individual. Lastly, $\mathcal{Y}$ describes outcome space for the observations of the outcomes of treatments *treatment_ outcomes.csv*. In our specific case, we have that $\mathcal{X} \subseteq [0,1]^{150}$ (after normalisation and standardisation), $\mathcal{A} = \{0,1\}^3$, and $\mathcal{Y} = \{0,1\}^8$.

In a first step, we wish to model the relation of the available data and other influences. We propose the following dependencies:

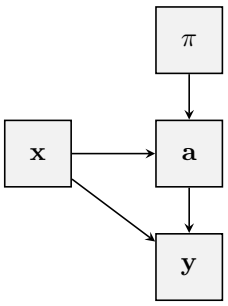

Figure 3: Model of the dependencies of the (unknown) decision-maker's policy $\pi(\mathbf{a} \mid \mathbf{x})$, feature vector $\mathbf{x} \in \mathcal{X}$, choice of treatment $\mathbf{a} \in \mathcal{A}$, and outcome $\mathbf{y} \in \mathcal{Y}$.

In Figure 3, we illustrate our understanding of the relation between features, treatments and outcomes. Here, we assume a underlying (unknown) policy $\pi : \mathcal{X} \to \mathcal{A}$ deployed by a decision-maker, mapping from features to treatments. While the policy π can play a crucial role, e.g. when a certain treatment is only prescribed to hopeless cases and thus the treatment shows close to

zero success rate, we will refrain from estimating π (thus implicitly assuming a uniform policy π).

To estimate the effect of treatments on alleviating symptoms, we will deploy a model

$$f : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{Y}),$$

i.e. a model mapping features and treatments to a probability distribution over outcomes $\Delta(\mathcal{Y})$, where $\Delta(\mathcal{Y})$ denotes the probability simplex of dimension $dim(Y) - 1$. More specifically, we will use Logistic Regression.

Our Approach. In order to verify the model, we will simulate data. The features are simulated as described in section 1A, with an addition of the feature Income, simulated by a half normal distribution with parameter beta=1. We simulate a situation where each individual has a probability for a symptom given no treatment, one treatment leads to a significant decrease of this probability, the other has a minor effect and the combination leads to some probability in the range between the two. Using this synthetic data, we run the following pipeline: 0) Balance the data (as described in 1A) 1) Train-test split of the data. 2) Train a logistic regression model and use 5-fold cross validation to select the hyperparameters. 3) Evaluate using the test set - we wish to see if the model successfully manages to discover this effect. In figure 4 we can see the difference in frequency of the symptom between the true simulated data and the predictions provided by the model. This difference is divided by the frequency of the symptom for scaling. This result is given for combinations of different effect sizes for the two treatments.

After the use of simulated data, we use the above pipeline on the given dataset. Once the model is trained, we can also provide additional synthetic samples to the model, with and without treatment, to see the effect of the treatment on samples with different features. We can see these examples in figure 5

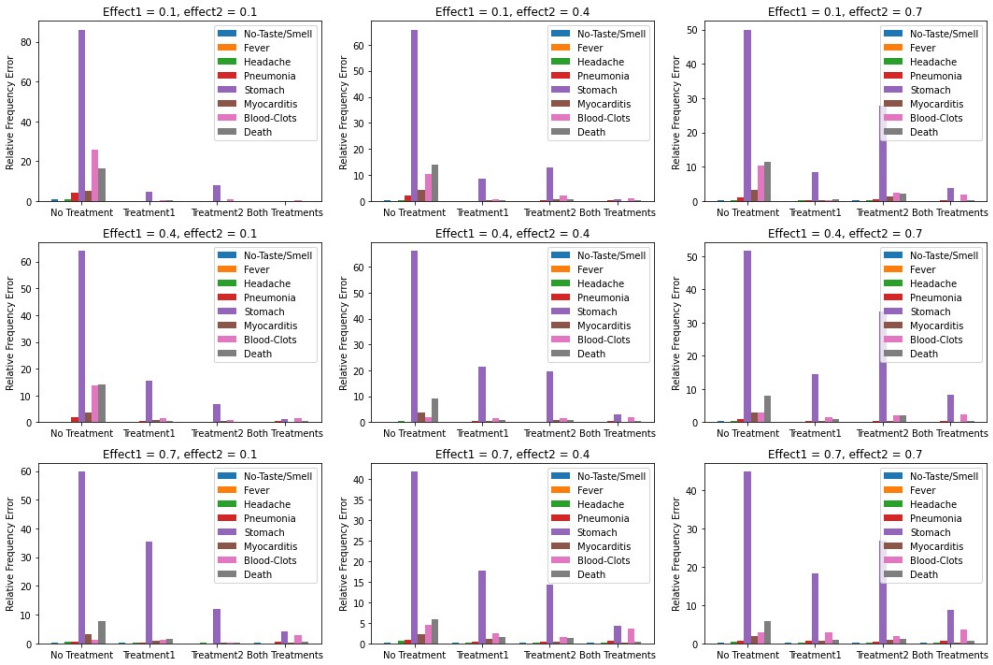

Figure 4: Difference in frequency of symptom between true simulated data and model predictions, divided by the frequency of the symptom for scaling, for different combinations of effect sizes for the two possible treatments.

3 Fairness and Privacy Considerations

In the context of clinical trials and testing of medications as well as vaccinations, fairness would guarantee that sufficiently many patients from different groups (e.g. age, race, or sex) are being included in the study so that medication (and vaccination) effectiveness and side-effects can be assessed correctly for all groups present in the population.

We clearly could have tried to create differentially private models. Surely, there is no guarantee of our current models (e.g. the model from Task 2) to be privacy-preserving. In particular, we use (and make available) a simulator based on 'real' data. Such a simulator that is based on real patient data ought to be privacy-preserving.

4 Code

The Colab Notebook can be found here: Link

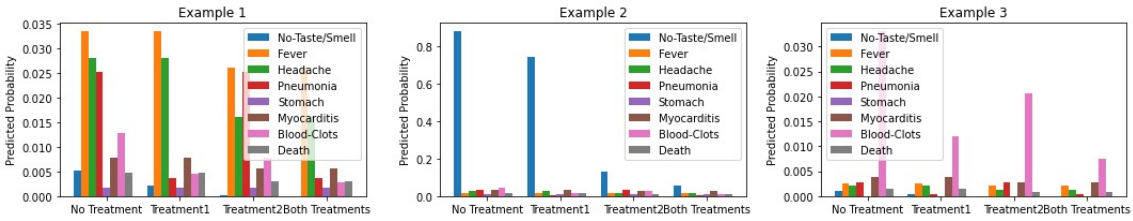

Figure 5: Symptom probabilities for 3 individuals, each assigned with no treatment, treatment1, treatment2 or both