IN-STK5000/9000 - Autumn 2021: Project1

Fábio Rodrigues Pereira
fabior@uio.no

Nicholas Walker
walker@nr.no

Aurora Poggi
aurorapo@uio.no

1 Introduction

At the time of writing this paper, Covid-19 is a current and serious threat to humans, which has killed $4.860.014$ out of $224.989.641$ [1] infected people around the globe. Given this, researchers have been assigned the duty of using intelligent ways to fight this illness. With access to patient data combined with considerable computing power, analysts and researchers can employ statistics, machine learning, and neural networks to find vital answers that can save people's lives.

In this paper, we propose and apply an experimental methodology to gain relevant knowledge in the understanding treatments and risk factors with Covid-19 illness. In particular, we investigate the population's age, gender, income, comorbidities, and any specific genome pattern predisposition to develop greater disease severity. Additionally, we analyze the efficacy of vaccination, its side effects, and the effectiveness of treatments to avoid death related to the Covid-19 illness. Finally, we will also discuss privacy issues that may arise with the data set.

In order to establish a ***ground truth*** and determine that our model is appropriate to the task, we generate synthetic data-bases. This will allow us to understand what information we can reliably extract via our model, i.e., basic feature correlations, probabilities, and descriptive statistics. With these considerations, we will be able to formulate a more precise approach to the following exact tasks:

1a. Questions:

   i. Can we predict death rate given age, gender, income, genes and comorbidities?
   ii. Which explanatory features are best to predict death?

1b. Questions:

   i. Can we predict death rate (efficacy) of a vaccine?
   ii. Which vaccine is most effective?

1c. Questions:

   i. Can we predict a specific symptom(s) (side-effect(s)) of a vaccine?
   ii. Which side-effect(s) each vaccine produce?

2. Questions:

   i. Can we predict death rate given a specific treatment?
   ii. Which treatment is the most effective?
   iii. Can we predict a precise symptom(s) (side-effect(s)) given a specific treatment?
   iv. Which side-effect(s) does each treatment produce?

To investigate each of these questions, we perform three automated methodologies. Our first methodology is to compute the autocorrelation matrix of the features of the data, in order to establish a high-level understanding of the data. The second approach is to evaluate the conditional probability of an outcome (e.g. $\mathbf{P}(Symptom|Vaccine)$ or $\mathbf{P}(Death|Vaccine)$). The third automated methodology is to model each question with a machine learning model, specifically *Logistic Regression*. To this end we develop a pipeline to train a logistic regression model on the data, starting again with a synthetic data set generated to establish the effectiveness of the model for modelling that question. Finally, we train a logistic regression model with cross-validation which then learns to predict an outcome based on input features.

[1] http://www.worldometers.info/coronavirus/

2 Reproducibility

This project can be reproducible by the code stored at our GitHub repository https://github.com/fabiorodp/IN_STK5000_ Adaptive_methods_for_data_based_decision_making/tree/main/project1. The results are achieved again with a high degree of reliability when the methodologies are replicated by the file ***main.py***.

3 Synthetic data-sets

A Python class called *Space* [2] is built to make the synthetic data-sets $(\Omega_1, \Omega_2, \Omega_3)$. We define the argument $seed = 1$ for maintaining consistency of the randomly generated values. For each Ω, there are $N = 100,000$ samples generated, denoting each individual. When treatment argument is false, there are 150 features in the data, but when *add_treatment=True*, 152 features with the following independent distributions:

[0] Covid_Recovered $\sim Binomial(1, 0.3)$

[1] Covid_Positive $\sim Binomial(1, 0.3)$

- Symptoms $\sim OneHotE(Uniform(1, 8))$

  [2] No-Taste/Smell
  [3] Fever
  [4] Headache
  [5] Pneumonia
  [6] Stomach
  [7] Myocarditis
  [8] Blood-Clots

[9] Death $\sim Binomial(1, 0.1)$

[10] Age $\sim Uniform(1, 100)$

[11] Gender $\sim Binomial(1, 0.5)$

[12] Income $\sim Normal(25000, 10000)$, where people with $age \leq 18$ have no income.

[13:141] Genes $\sim Binomial(1, 0.25)$

[141] Asthma $\sim Binomial(1, 0.07)$

[142] Obesity $\sim Binomial(1, 0.13)$

[143] Smoking $\sim Binomial(1, 0.19)$

[144] Diabetes $\sim Binomial(1, 0.10)$

[145] Heart-disease $\sim Binomial(1, 0.1)$

[146] Hypertension $\sim Binomial(1, 0.17)$

- Vaccines $\sim OneHotE(Uniform(1, 4))$

  [147] Vaccine 1
  [148] Vaccine 2
  [149] Vaccine 3

[150] Treatment 1 $\sim Binomial(1, 0.7)$

[151] Treatment 2 $\sim Binomial(1, 0.5)$

It is essential to state that a *pandas.DataFrame* object called *Space*, inside the class Space, is generated automatically by the constructor of that class.

Next, the Space's class method *assign_corr_death()* is called to assign new values for the feature $Death$, from a *Binomial* distribution, based on a pre-defined combination of conditional probabilities between the explanatory variables (Age, $Income$, $Diabetes$, $Hypertension$, $Gene_1 + Gene_2$, $Vaccine1$, $Vaccine2$, $Vaccine3$, $Treatment1$ and $Treatment2$) with the response variable ($Death$). These probabilities apply only for the cases when $CovidPositive$ is true.

Now we need some other conditional probabilities to differentiate each Ω used for each specific task of this project. Therefore, in order to answer questions in 1a, Ω_1 requires features $No\_Taste/Smell$, and $Pneumonia$ drawn from independent Binomial distributions with the following conditional probabilities:

$$\mathbb{P}(No\_Taste/Smell|CovidPositive) = 0.8,$$

$$\mathbb{P}(Pneumonia|CovidPositive) = 0.5.$$

For the questions in tasks 1b and 1c, we will use Ω_2 with three other re-generated features $Blood\_Clots$, $Headache$, and $Fever$ which are drawn from independent Binomial distributions with the following conditional probabilities:

$$\mathbb{P}(Blood\_Clots|Vaccine1) = 0.3,$$

$$\mathbb{P}(Headache|Vaccine2) = 0.6,$$

$$\mathbb{P}(Fever|Vaccine3) = 0.7.$$

For the questions in task 2, two new features will be introduced, $Treatment1$ and $Treatment2$, which can be generated when calling the Space class with argument *add_treatment = True*. Then,

[2] https://github.com/fabiorodp/IN_STK5000_Adaptive_methods_for_data_based_decision_making/blob/main/project1/helper/generate_data.py

the treatments will be pulled from Binomial distribution with probability equals 70% and 50% for each treatment respectively. Note that both treatments can occur for the same individual. After that, we will re-generate values for $Death$ by using the class method *assign_corr_death()*, and when calling the class method *add_correlated_symptom_with()*, we will assign new values for the features $Headache$ and $Fever$ from independent Binomial distributions with conditional probabilities as follows:

$$\mathbb{P}(Headache|Treatment1) = 0.5,$$

$$\mathbb{P}(Fever|Treatment2) = 0.7.$$

After generating synthetic spaces $(\Omega_1; \Omega_2; \Omega_3)$ for each task (1a; 1b and c; 2) respectively, with all features properly designed and correlated, we can move forward with our study to understand the ground truth (i.e. effectiveness of our analysis) and estimate parameters on observational and treatment data.

4 Methodologies

In **Methodology**$_1$ we compute the autocorrelation matrix, also called series correlation, see (Devore and Berk), between features to get a first view of the available data and the relationships between them. To do this we use the autocorrelation matrix that explains the degree of dependence between the values of our features. Given a sequence x, the autocorrelation is given by

$$R_{xi} = Corr(x, x)_i = \sum_{i=0}^{\infty} x_i x_{i+1}.$$

In this project we did not use this formula but simply applied the .corr() function of Pandas.

In **Methodology**$_2$, we calculate the conditional probability of a response variable given an explanatory variable, $\frac{P(A \cup B)}{P(B)}$, where deeper explanation can be found at (Walsh, 2011). This calculation assumes only a relationship between the single explanatory variable and the response and therefore we note that this calculation does *not* model joint probability distributions, because the number of features makes it infeasible to model all combinations of them. While this assumption is not true in practice (see discussion of methodology 1 in Section 6 for correlation between features), we can nonetheless analyze the likelihood of a given outcome from the perspective of only knowing a single feature (e.g. the chance of death given that one has pneumonia). To give a bound of estimates, we perform this calculation on samples of our dataset, using a sample of a size $N = \frac{Length(Dataset)}{4}$. The data is sampled with replacement ("bootstrapped") and the proability is calculated on this sample of the data. We repeat this process 1000 times, and observe the quantiles of the resulting probability estimates to characterize the effect of each explanatory variable on the response. This process is repeated for synthetic, observational, and treatment data, and the details of the calculation $\frac{P(A \cup B)}{P(B)}$ is performed using the Pandas library (see the github repository for details of the groupby and division functions).

In **Methodology**$_3$ we use a logistic regression model to estimate the probability that the input data ($\mathbf{x}$) leads to outcomes (y_i), for $i = 1, 2, ..., m$, where m is the number of possible labels for the outcome classes and $y_i \in [0, 1]$. In the data we use, an outcome of 1 corresponds to death, and 0 to survival. The probability of the two outcomes is then as follows:

$$\mathbb{P}(y_i = 1|x_i, \beta) = \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}$$

$$\mathbb{P}(y_i = 0|x_i, \beta) = 1 - \mathbb{P}(y_i = 1|x_i, \beta),$$

where β are the coefficients that will be estimated by the model, see (Géron, 2017).

We also apply a regularization parameter $\lambda \geq 0$, which scales the model parameters. Then, the cost function of the model will be:

$$C(\boldsymbol{\beta}) = (\boldsymbol{z} - \boldsymbol{X\beta})^T(\boldsymbol{z} - \boldsymbol{X\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}$$

Where $\mathbf{X}$ is the input data and z is the output (true) labels.

We train the logistic regression models on a balanced set of data points sampled from the full data, in order to ensure that the classifier properly learns to predict each class (rather than e.g. predicting only $Death = 0$ because the vast majority of patients survive). After this selection, we define a pipeline for feature selection and hyper-parameter tuning which is applied in randomized-search cross-validation ($CV = 500$). This is to ensure that the model parameters are accurate and that we obtain a range of values with which to estimate them.

The logistic regression models will then learn coefficient values from the data corresponding to each feature, where higher coefficient values are predictive of death and negative coefficient values are predictive of survival. Lastly, we select the features with the highest and lowest (i.e. greatest absolute value) coefficients to determine which of them are most predictive of symptoms, for both vaccine side-effects as well as treatment effectiveness.

We implement the model pipeline, logistic regression model itself, and the cross-validation using the `statsmodels` package. This is convenient because the library provides functionality to display a model summary, including summary statistics for coefficient values including mean, median, standard deviation error, p-values, and confidence intervals (the last of which we detail in Section 6). The full printout of these results can be obtained by running the code contained in the GitHub repository.

5 Baseline

In this section, we analyze the results obtained from applying our methodologies on synthetic data and establish their efficacy.

5.1 Ground truth: Task 1_a

For task1_a, we use Ω_1 as our synthetic data and calculate the 10 most negatively correlated and 10 most positively correlated features with $Death$.

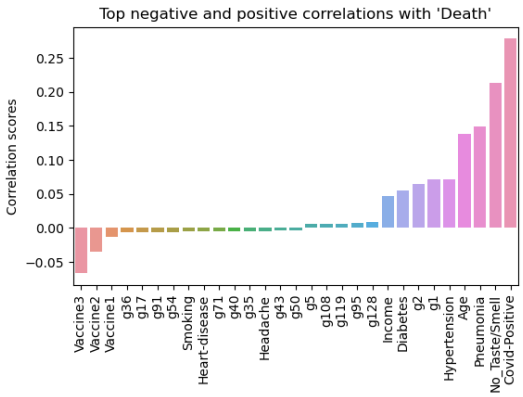

Figure 1: Correlations between Death and other features for the set Ω_1.

In Figure 1 we look at the correlation between $Death$ and some of our features. As expected, vaccines are the most negatively correlated with $Death$, especially $Vaccine3$, which seems to be the most efficient. On the other hand, $CovidPositive$ appears most predictive of death, followed by some symptoms.

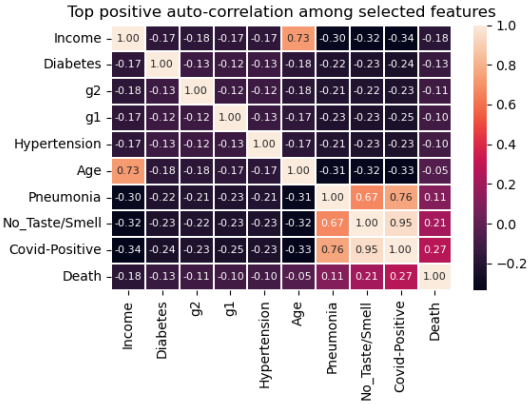

Figure 2: Auto-correlations among features for the set Ω_1.

In Figure 2, we observe the auto-correlations between features. We notice a high correlation between $CovidPositive$ and some symptoms, such as $No\_Taste/Smell$ and $Pneumonia$. In addition, there is a high auto-correlation between $Income$ and Age. For this reason, we consider only the $CovidPositive$ population and remove the $Income$ variable from this data to obtain all independent explanatory variables, which is a necessary assumption for our chosen $Logistic\ Regression$ model.

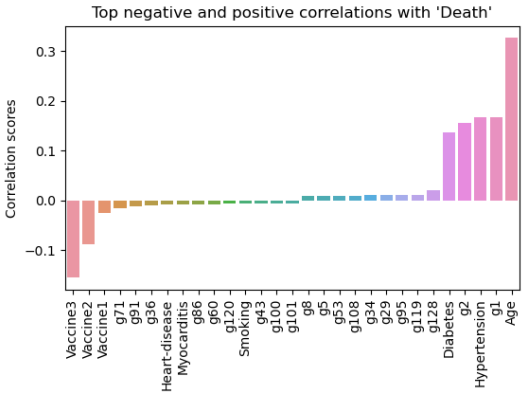

Figure 3: Correlations between $Death$ and selected features for the set Ω_1.

At this point, we observe the new **Figure 3** containing a distinct correlation-rank. Indeed, there is a significant difference from **Figure 1**, because previously the feature $CovidPositive$

brought some auto-correlated symptoms to the top-correlated features with $Death$. This is to say that, since $CovidPositive$ is highly correlated to $Death$ and some symptoms are auto-correlated to $CovidPositive$, then these symptoms seemed to also be correlated with $Death$. However, this logic contradicts the underlying process to generate the synthetic data and is not desirable. When we analyze the data within the $CovidPositive$ population, we can then notice that these specific symptoms do not explain $Death$, but only $CovidPositive$. Under this analysis, the auto-correlated symptoms features disappear from the top-ranked and are replaced by other variables that actually explain $Death$ according to our design of the synthetic data.

Our results here establish that methodology$_1$ can reliably identify features which affect patient outcomes either positively or negatively. The method reliably and correctly identifies the known positive relationship between the Age, $Gene_1$, $Hypertension$, $Gene_2$, $Diabetes$ features with $Death$ as we defined in the synthetic data generation.

Following methodology$_1$, we then apply methodology$_2$ on Ω_1.

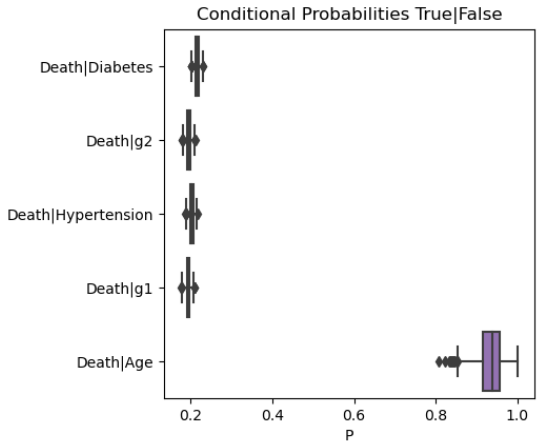

Figure 4: Conditional Probability of Death in Ω_1 without comorbidities.

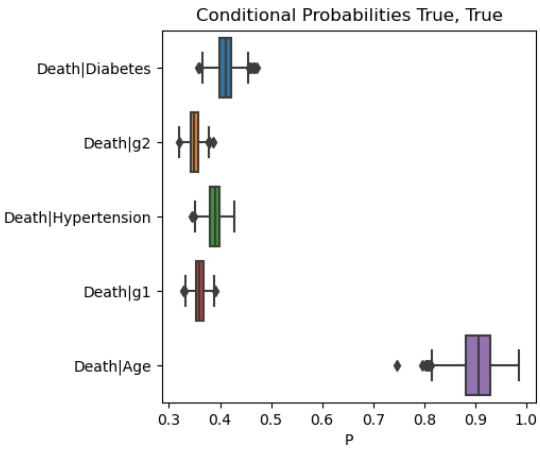

Figure 5: Conditional Probability of Death in Ω_1 with comorbidities.

In **Figures 4** and **5** we observe the conditional probability when comorbidities are not present and present, respectively. The conditional probability of death is observed to be higher when $Diabetes$ and $Hypertension$ are present, as well as with $Gene_1$ and $Gene_2$. In the latter case, the effect appears less pronounced, however in our synthetic data it is the combined presence of both genes that increases the probability of death, so individually their impact on the chance of death is less noticeable in this methodology. Nonetheless, the features which cause death in our synthetic data are correctly identified and the increased effect on death is clearly shown over the 1000 probability estimates.

Finally we apply methodology$_3$ to Ω_1.

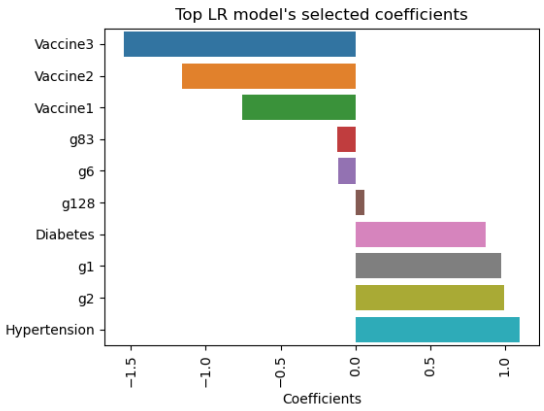

Figure 6: Coefficient Values of Logistic Regression on Ω_1, after randomized grid-search 500-fold CV.

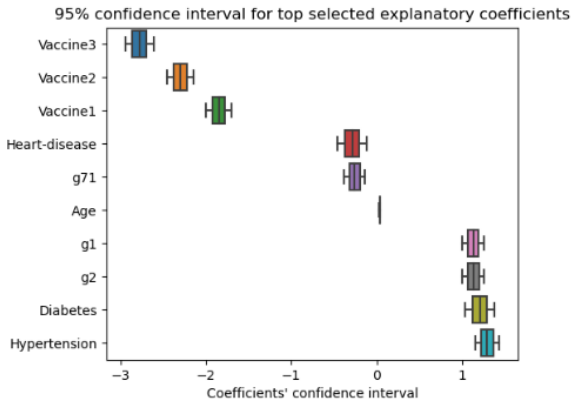

Figure 7: 95% Confidence Interval for Logistic Regression Coefficient Values.

As described in Section 4, we train a *Logistic regression* classifier to predict $Death$. The best performing model scored an accuracy of 0.82, and we then observed the coefficient values of the model. In **Figure 6**, we show the 5 largest positive and 5 largest negative coefficient values for the classifier. The vaccines have by far the largest negative coefficient value in predicting $Death$, as expected. On the other side, $Hypertension$, $Gene_2$, $Gene_1$, and $Diabetes$ are the most useful features for the model to predict $Death$. We can then conclude that as with the first two methodologies, methodology$_3$ accurately identifies the features which predict death and our overall framework for analyzing the real data is robust.

5.2 Ground truth: Task $1_b, 1_c$

In this subsection we use another synthetic dataset Ω_2, in which we imposed a correlation between each vaccine and a specific side effect.

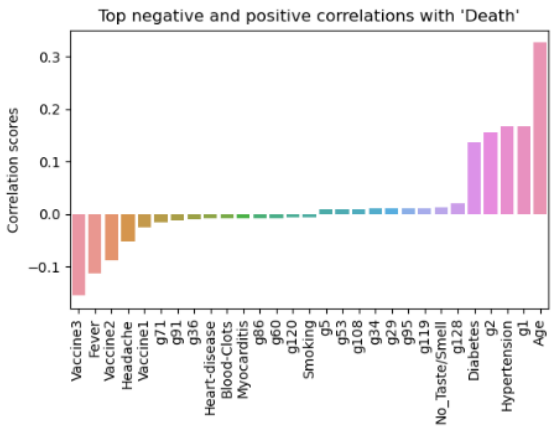

Figure 8: Correlations between Death and selected features for the set Ω_2.

Similarly to **Figure 1**, in **Figure 8** we observe the positive and negative correlations between features and death. These two figures are very similar, with the exception that side-effects from the vaccines now appear among negatively correlated features. These results are consistent with the generation of Ω_2 as described in Section 3.

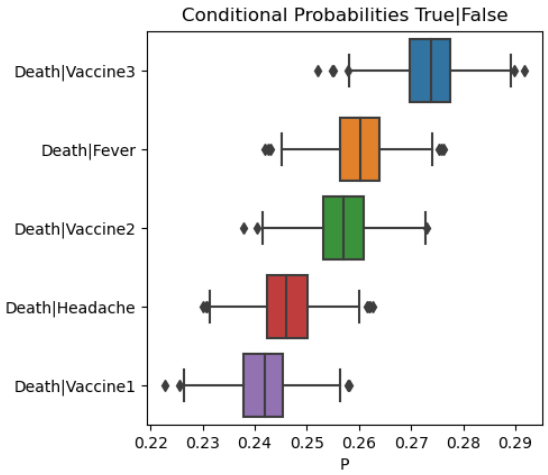

Figure 9: Conditional Probability of Death in Ω_2 without comorbidities.

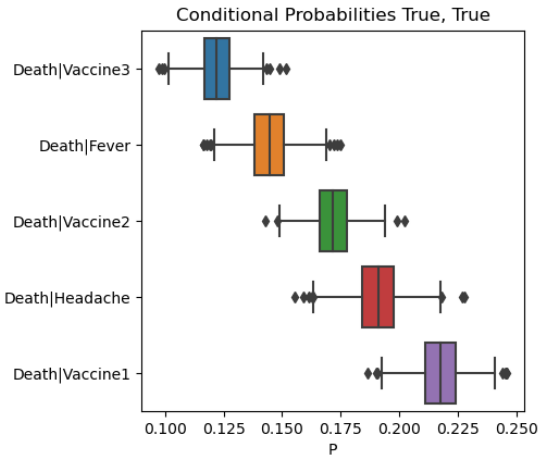

Figure 10: Conditional Probability of Death in Ω_2 with comorbidities.

Figures 9 and **10** again show the effect of the vaccines in reducing the likelihood of death, as defined for Ω_2. We then also investigate the conditional probability of symptoms given vaccines.

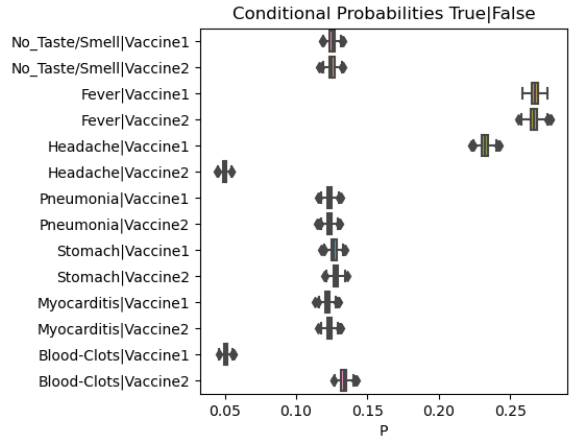

Figure 11: Conditional Probability of Symptoms in Ω_2 without comorbidities.

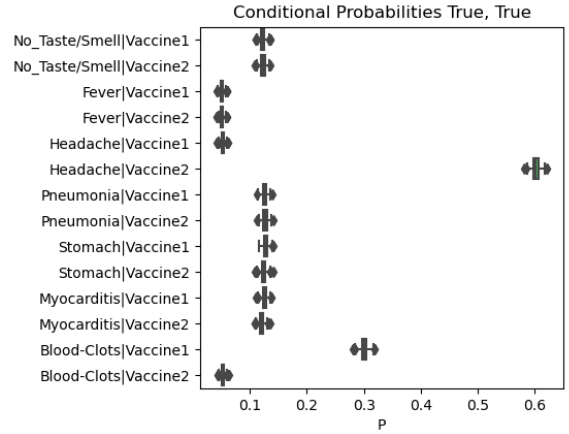

Figure 12: Conditional Probability of Symptoms in Ω_2 with comorbidities.

These plots show the side-effect we defined for two of the vaccines (omitting $Vaccine3$ for purposes of space) in Ω_2. The chance of having a $Headache$ is accurately shown to be affected by $Vaccine2$, and likewise the relationship between $Blood\_Clot$ and $Vaccine1$.

Lastly, we analyze Ω_2 with our third methodology. As in Task $1a_{5.1}$, the model is able to learn the correct features to predict $Death$, which are the same as before because we are training the model to predict the same outcome. We can then conclude that our methods analyze the Ω_2 synthetic data appropriately.

5.3 Ground truth: Task$_2$

For task$_2$, we use Ω_3 as our synthetic data. In this subsection we are interested in the effect of treatments on alleviating symptoms and preventing deaths.

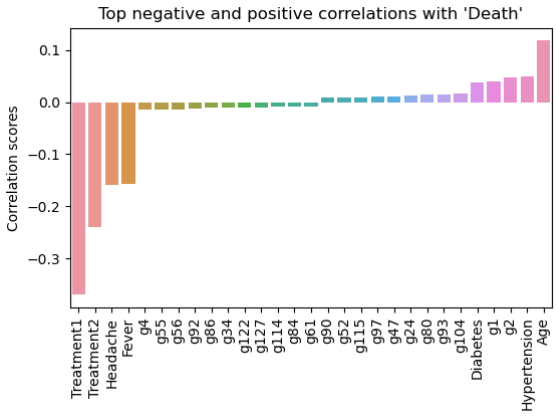

Figure 13: Correlations between $Death$ and other features for Ω_3.

In **Figure 13** we examine the top negative and positive correlations with $Death$. As expected, among the top negative features we find the two treatments while in the positive features we observe Age, $Hypertension$, $Gene_1$, $Gene_2$ and $Diabetes$. We defined the data such that both treatments are effective, with $Treatment1$ being more effective (thus both are highly negatively correlated to death). Additionally, $Treatment1$ is associated with a side effect of $Headache$ and $Treatment2$ is associated with a side effect of $Fever$ and these features are thereby negatively correlated with $Death$ due to their auto-correlation with the treatments.

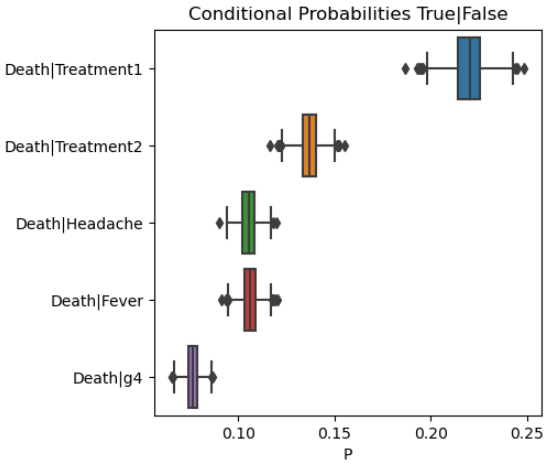

Figure 14: Conditional Probability of Death in Ω_3 without comorbidities.

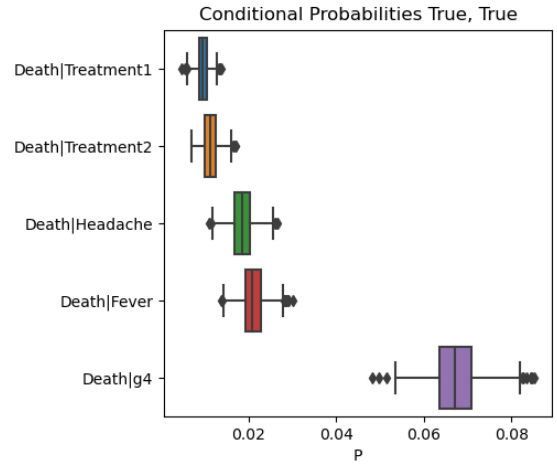

Figure 15: Conditional Probability of Death in Ω_3 with comorbidities.

Methodology 2 as applied to Ω_3 shows the influence of treatment on the chance of death. When treatments are not applied, the chance of death is above 14% as in **Figure 14**, both treatments are given, the probability of death is reduced to 0.01 as in **Figure 15**. Each treatment also individually reduces the chance of death. **Figure 16** shows the probability of symptoms after treatment has been applied. Treatments predict each other's side effects because it is possible for a patient to be treated with both.

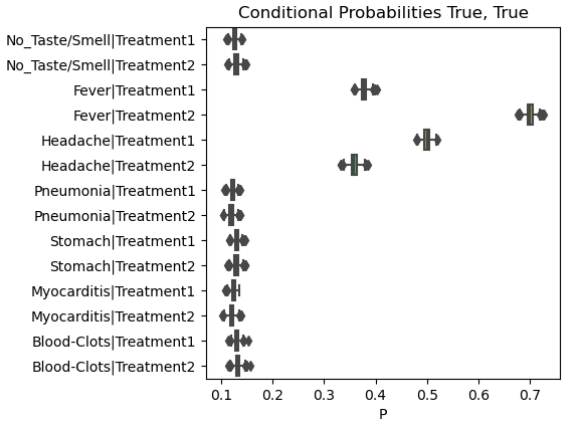

Figure 16: Conditional Probability of Symptoms Given Treatments for Ω_3.

Our final test on synthetic data is applying methodology$_3$ to Ω_3. The methodology here returns the coefficient values for a *Logistic regression* model predicting death given treatment.

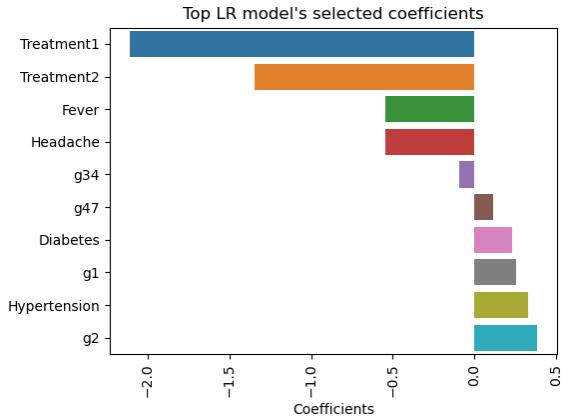

Figure 17: Coefficient Values of Logistic Regression on Ω_3, after randomized grid-search 500-fold CV.

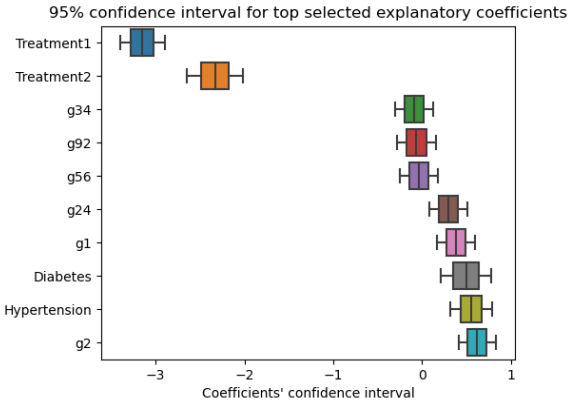

Figure 18: 95% Confidence Interval for Logistic Regression Coefficient Values.

In Figure 17 we show the top positive and negative coefficient values for this model. The best model accuracy for this data was 0.86. These coefficients accurately show that treatments are highly predictive of survival in Ω_3. **Figure 18** shows the confidence interval for the coefficient values, showing that the model training procedure results in a useful range of values that allow the correct conclusion to be drawn that treatments are effective.

Therefore, given all of the experiments performed on the three synthetic datasets, we are able to determine that each of our methodologies are capable of inferring the correct conclusions to explaining the synthetic data and match how we created this data.

6 Results

In this section we describe the results of our methodologies on the observational and treatment data. We investigate the relations between features

and symptoms of the real data provided in the files[3] and determine answers to our questions from the data.

6.1 Observational Data: Task 1_a

In this subsection we use the real data provided in the *Observation-feature.csv* file.

Applying methodology$_1$ to the observational data, we obtain the following correlations between features and $Death$ shown in Figure **Figure 19**. We firstly observe that vaccines have relevant negative correlations, while for positive correlations $CovidPositive$ is the major influencer of $Death$.

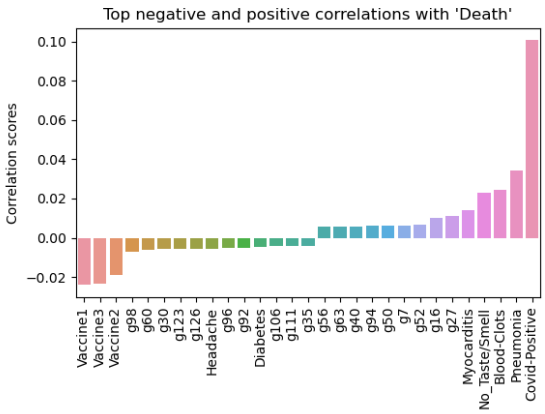

Figure 19: Correlations between Death and other features for Observational Data.

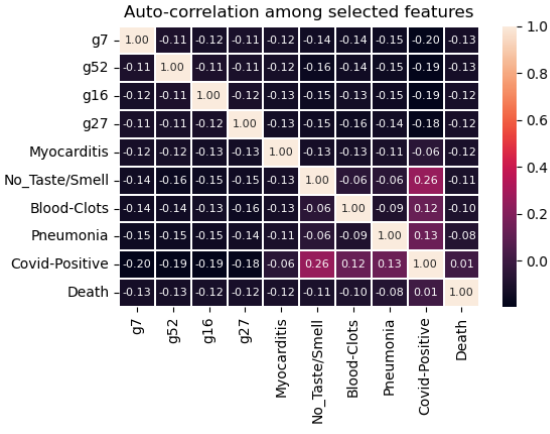

Figure 20: Auto-correlations among features for Observational Data.

The auto-correlation matrix in **Figure 20** shows high values of correlation between $CovidPositive$ and some symptoms: such as $No\_Taste/Smell$, $Pneumonia$ and $Blood\_Clots$. Hence, similarly to the synthetic data we found that it is necessary to work in a subset of the data containing only the $CovidPositive$ individuals to eliminate the dependence of symptoms on it. As our experiments with synthetic showed, this should allow us to determine the features which are actually predictive of death.

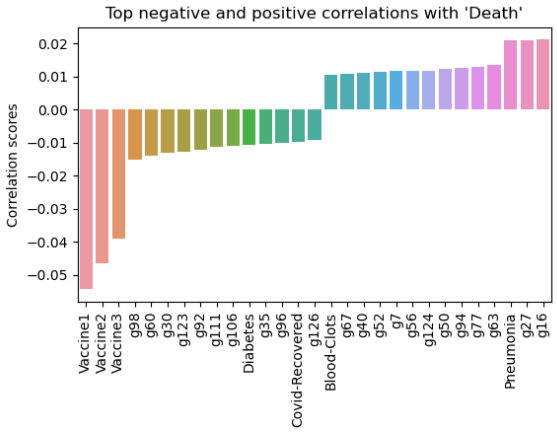

Figure 21: Correlations between Death and selected features for Observational Data.

Then we obtain the following correlations, where we can observe that vaccines are much more strongly predictive of survival. Perhaps even more notably, the features most predictive of death are no longer dominated solely by symptoms, but now show several genes which are especially predictive of death, namely $Gene_{16}$, $Gene_{27}$, $Gene_{63}$, $Gene_{77}$ in the top 5 features. The only symptom remaining among the top 5 predictive features was $Pneumonia$.

Following from the useful indications from methodology$_1$, we then use methodology$_2$, yielding the conditional probability of death given explanatory features.

[3] See GitHub

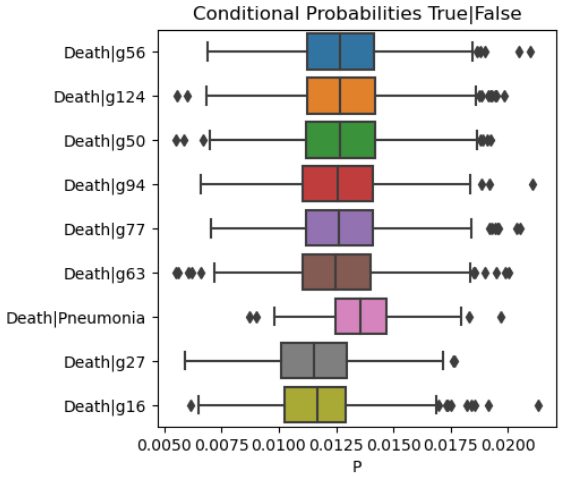

Figure 22: Conditional Probability of Death given $Feature = 0$.

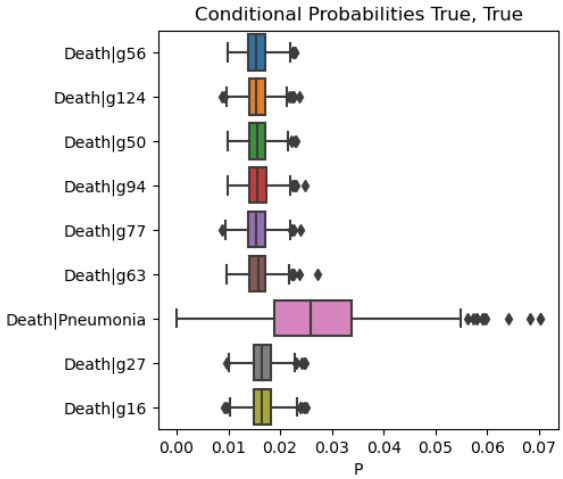

Figure 23: Conditional Probability of Death given $Feature = 1$.

Figures 22 and **23** show the conditional probability of death given the features from the dataset. We see in the plots that the probability of death increased by the greatest amount when conditioned on $Pneumonia$, from a mean of approximately 0.013 to 0.025. We also see a small increase in the probability of death when conditioned on the gene features, however the range of values when the genes are present versus not present overlap quite heavily. In this regard, we are not able to deduce that these genes affect the probability of death from methodology$_2$.

We then apply our final methodology to the observational data to determine our final conclusions with respect to the effect of genes and symptoms on death in the observational data.

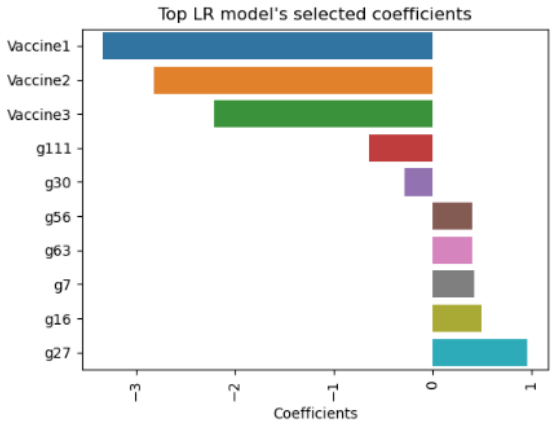

Figure 24: Coefficient Values of Logistic Regression observational data, after randomized grid-search 500-fold CV.

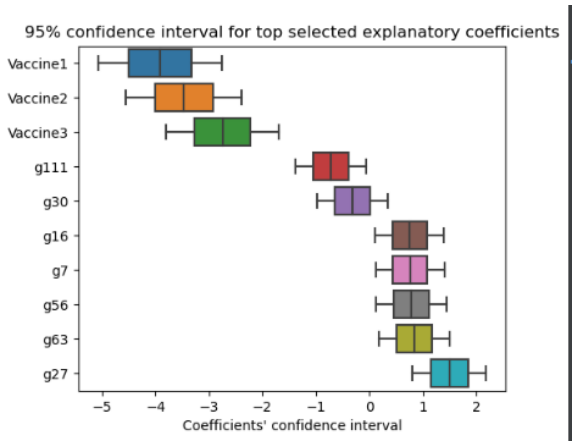

Figure 25: 95% Confidence Interval for Logistic Regression Coefficient Values.

Following our procedure, the *Logistic regression* models were trained on the top 10 negatively correlated and top 10 positively correlated features from observational data. When trained with all features, including $CovidPositive$, the logistic regression models were unable to predict $Death$ at any better than 0.5 (i.e. random guess). When the models were trained on a subset containing only the $CovidPositive$ population without feature selection, the best model had an accuracy of 0.71. Finally, when using the 10 most negatively and 10 most positively correlated features, the best scoring model had an accuracy of 0.87. This approach therefore remains the same in our subsequent tasks and models trained for methodology$_3$.

The resulting coefficient values are as shown in 24, with 95% confidence intervals for the logistic

regression shown in **Figure 25**. This is to say that we are 95% confident that the true value of the coefficient will be within the intervals shown in the figure. From this, we can conclude that if a feature has a lower range of possible values, then it is less predictive of death (or more predictive of death if the range is higher than other features).

Based on these results along with the previous methodologies, we return to our first set of questions:

- *Can we predict death rate given age, gender, income, genes and comorbidities?* The logistic regression coefficients and accuracy show that it is possible to train a model which is able to accurately predict death.

- Our second question (*Which explanatory features are best to predict death?*) is answered by these results as well, showing that the symptoms and comorbidities which are most effective in predicting death are $Gene_{16}$, $Gene_{63}$ and $Gene_{27}$. $Pneumonia$ did not arise in methodology$_3$ as a top predictor of death, and $Gene_7$ does not appear from methodologies 1 and 2, so we conclude that these features are most likely comparatively less important predictors of death (though likely still risk factors).

6.2 Observational Data: Task $1_b, 1_c$

In this section we apply our methodologies for estimating the efficacy of vaccines and investigate their possible side-effects. The results of methodology 1 are the same as from Subsection 6.1, but in this case we focus on the fact that each of the vaccines are highly negatively correlated to death (**Figure 21**), indicating that they are possibly effective in preventing it.

On the other hand, methodology$_2$ differs from our previous analysis, as in this case we are modelling the conditional probability of side-effects (including death) from vaccines.

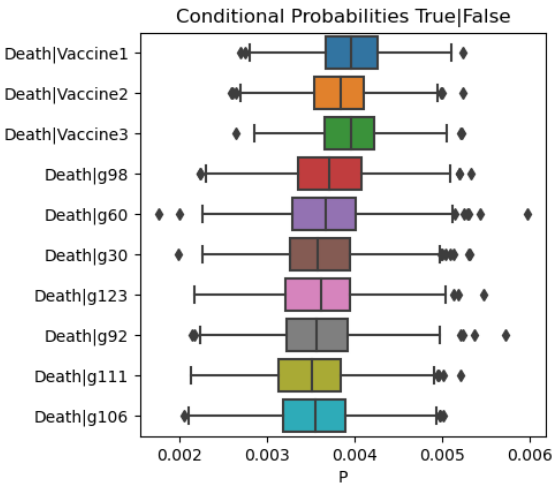

Figure 26: Conditional Probability of Death given $Feature = 0$.

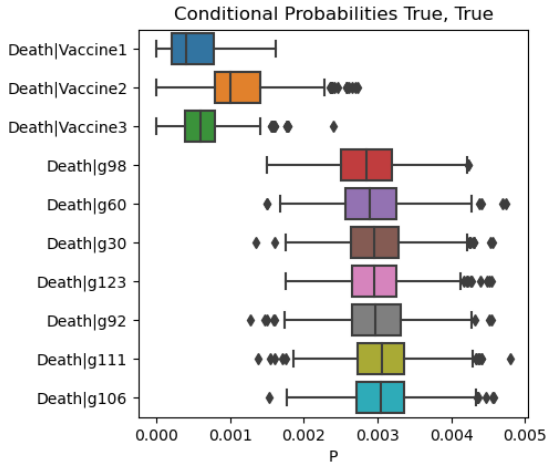

Figure 27: Conditional Probability of Death given $Feature = 1$.

Figures 26 and **27** show the conditional probabilities of side-effects where the patient is unvaccinated and vaccinated, respectively. The figures show that the probability of $Death$ is greatly reduced when conditioned on all of the vaccines, with $Vaccine1$ being the most effective (specifically, reducing the probability of death the most). In this case we also observe that the quantiles for the probability of death given vaccines are greatly below the probability of death without vaccines. Interestingly, while both this methodology and methodology$_1$ show $Vaccine1$ to have the greatest effect on reducing the probability of death, they differ on which of $Vaccine2$ or $Vaccine3$ is the second most effective in preventing death.

An additional aspect of our analysis with methodology$_2$ is the estimation of side effects from

vaccines. With this methodology we are also able to calculate the conditional probability of symptoms given vaccines. **Figures 28** and **29** show that the probability of $Fever$ and $Headache$ symptoms was greatly increased when conditioned on the vaccines. Without vaccines the mean probability of $Headache$ was 0.03 and $Fever$ was 0.05. 0.05 for headache, 0.085 for Fever. We did not find a similar association with any other side-effects.

The mean probability of $Headache$ symptom increased from approximately 0.03 to 0.05 when conditioned on each of the vaccines (all resulted in similar mean values), while the mean probability of $Fever$ increased from 0.05 to 0.085. For the probability of $Fever$ and $Headache$ conditioned on each of the vaccines, the quantiles of the probability estimate did not overlap with the probability estimate conditioned on not receiving the vaccine. For this reason we conclude that this is a reasonable indication that the vaccines are associated with $Fever$ and $Headache$ as side effects.

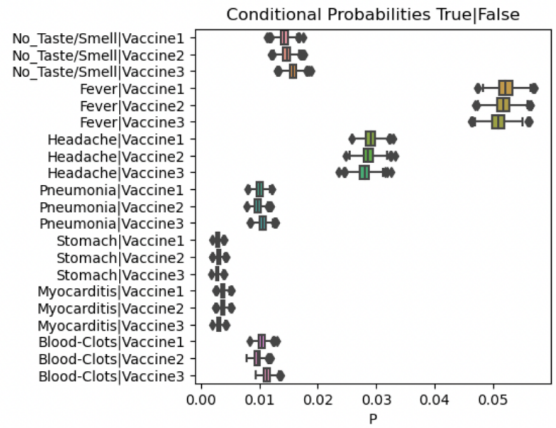

Figure 28: Conditional Probability of Symptoms given $Vaccines = 0$.

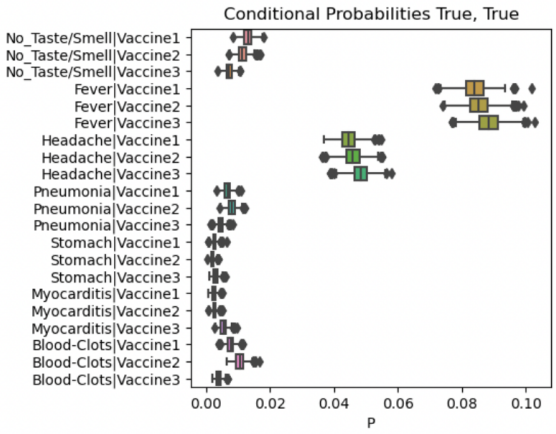

Figure 29: Conditional Probability of Symptoms given $Vaccines = 1$.

Next, turning to methodology$_3$, we train $Logistic\,regression$ models predicting symptoms including death. As with methodology$_1$, we refer to the same models described in 6.1 to determine the answer to our second set of questions (Tasks 1b and 1c). The top 3 negative features in the logistic regression model are the three vaccines (**Figure** 24), indicating that they are the most predictive of survival. The order of greatest effectiveness can be loosely observed as $Vaccine1$, $Vaccine2$, $Vaccine3$. This Nonetheless, the confidence intervals on the effectiveness of each vaccine (as measured with the coefficient values) broadly overlap (**Figure 25**) and all three can be considered "effective" under this measure (and there may be no reason for a patient to, for instance, prefer $Vaccine1$ to $Vaccine3$ in practice).

In conclusion with this, we answer the questions posed for Tasks 1b and 1c.

- *Can we predict death rate (efficacy) of a vaccine?* Yes, our methodologies indicate that all three vaccines are very effective in reducing the chance of death in the Covid-Positive population.

- *Which vaccine is most effective?* Our analysis indicates that $Vaccine1$ is the most effective of the three. However, we cannot definitively say whether $Vaccine2$ or $Vaccine3$ is superior, as the results of our methodologies differ on which is more predictive of survival.

- *Can we predict a specific symptom(s) (side-effect(s)) of a vaccine?* Through analyzing the conditional probability estimates obtained in

methodology$_2$, we predict that all three of the vaccines in the data are associated with side-effects of $Headache$ and $Fever$. In the case of features such as blood clots, the conditional probability estimates do not differ sufficiently from the conditional probability of death without the vaccines to conclude that the vaccines may cause them as side effects. Blood clots may be associated with $Vaccine2$ as shown in 29, but the overlap between this estimate and the baseline is too great to rule out the possibility of this being merely random chance.

- *Which side-effect(s) each vaccine produce?* From our analysis it appears that all three vaccines produce $Fever$ and $Headache$ as side-effects, with the unconfirmed possibility of blood clots from $Vaccine3$, as noted above.

6.3 Treatment Data: Task 2

In the final subsection we use the real data contained in the files 'treatment_features.csv', 'treatment_action.csv' and 'treatment_outcome.csv' regarding the treatments.

We again perform methodology$_1$ in order to obtain the correlations of features in the data. The results shown in **Figure 30** are similar to those in the observational data, with the new addition of $Treatment2$ as a variable which is highly negatively correlated with death. Notably, we do not observe $Treatment1$ in the top 10 negatively correlated variables.

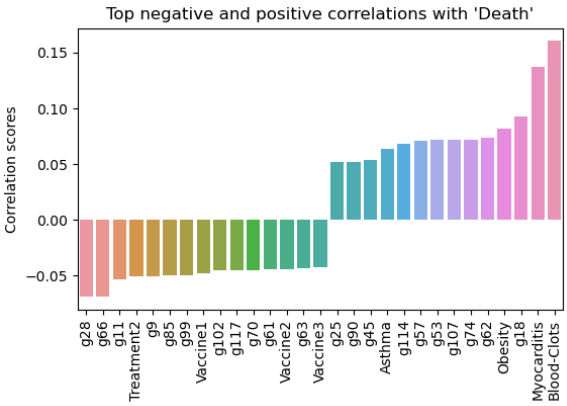

Figure 30: Correlations between Death and selected features for Observational Data.

We then proceed with analyzing **Figures 31** and **Figures 32** using methodology$_2$. The results show that the probability of $Death$ conditioned on $Treatment2$ appears to be lowered, but there is some overlap in the estimates from this method. For this reason, these results may not be as indicative in establishing that $Treatment2$ is effective in preventing $Death$.

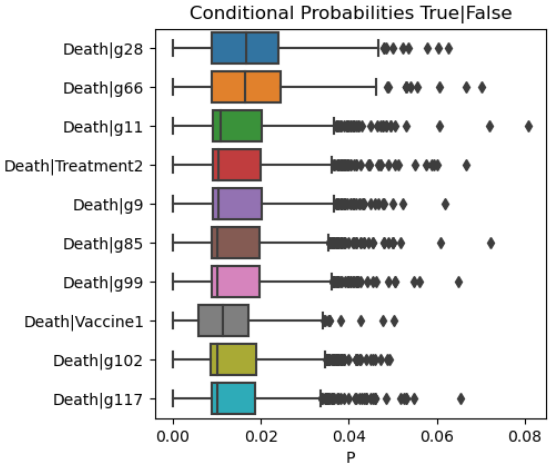

Figure 31: Conditional Probability of $Death$ given selected $Features = 0$.

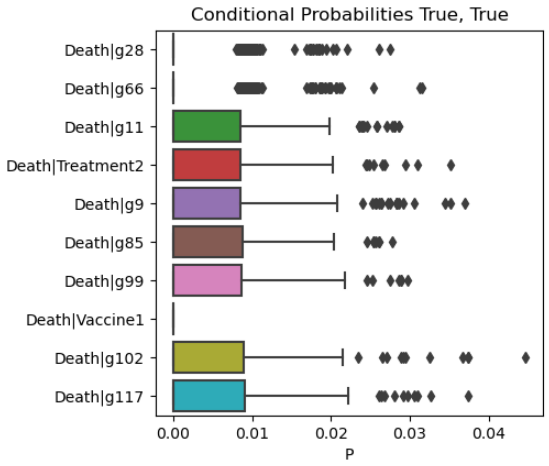

Figure 32: Conditional Probability of $Death$ given selected $Features = 1$.

We also observe in **Figures 33** and **34** that the probability of symptoms when conditioned on treatments are very similar to the probability when conditioned on the absence of treatment. The one notable feature of these charts is that the probability of $Blood\_clots$ conditioned on $Treatment1$ is 0. This is however a very marginal result and so we can not establish any evidence of side-effects from treatments using this method. We note that the amount of total samples in the data was very small, particularly with respect to certain symptoms. This may be a reason that this methodology

does not reach a clear conclusion and estimates several probabilities precisely at 0.

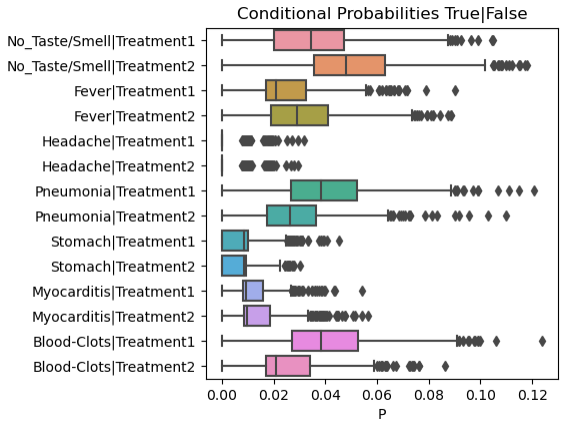

Figure 33: Conditional Probability of Symptoms given $Treatments = 0$.

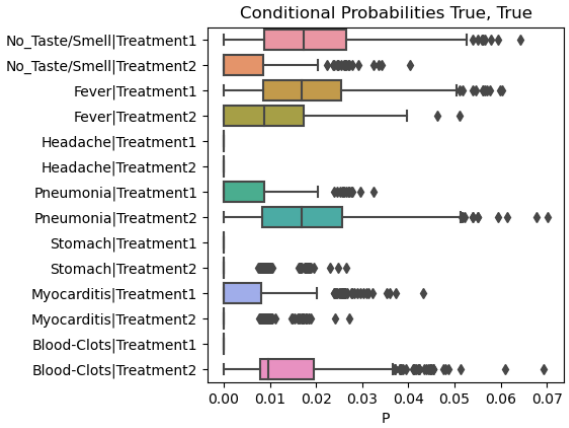

Figure 34: Conditional Probability of Symptoms given $Treatments = 1$.

Our last analysis of the treatment data is performed with methodology$_3$. In this case, due to the scarcity of data points, we train the $Logistic\ regression$ model using bootstrapping to resample the data, in order to balance the outcomes for the training process. The resulting best model accuracy is 0.98 with confidence intervals as in **Figure 36**. We observe once again in **Figure 35** that the the coefficient for $Treatment1$ is not in the top selected features. On the other hand, $Treatment2$ is once again among the most negatively associated variables with $Death$. Combined with our observation from methodology$_1$, this indicates that $Treatment2$ is effective at preventing $Death$, and very likely more effective than $Treatment1$.

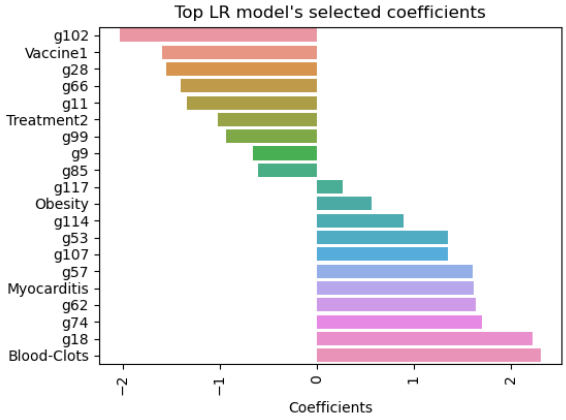

Figure 35: Coefficient Values of Logistic Regression, after randomized grid-search 500-fold CV.

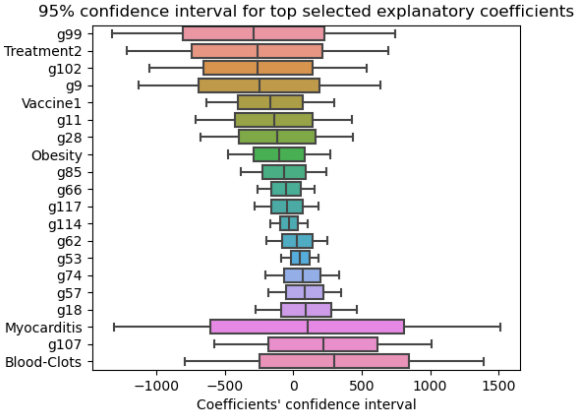

Figure 36: 95% Confidence Intervals for Logistic Regression Coefficient Values.

We are then able to answer the questions we posed for Task 2.

- *Can we predict death rate given a specific treatment?* Yes, we are able to train a logistic regression model which is able to predict death among the $CovidPositive$ population with an cross-validated accuracy of 0.98.
- *Which treatment is the most effective?* We estimate that the probability of death for a patient given $Treatment2$ is lower than without treatment, as well as lower than when given $Treatment1$. While $Treatment1$ may still be effective in preventing death (we do not establish that it is not), our analysis indicates that $Treatment2$ is more effective.
- *Can we predict a precise symptom(s)(side-effect(s)) given a specific treatment?* We were not able to establish that any side-effects were

more likely given either of the treatments. We speculate that this may be due to the low number of samples for each symptom, such that our calculation of conditional probabilities was not able to make an precise estimate approximating the actual outcomes.

- *Which side-effect(s) does each treatment produce?* As noted in the previous answer, we cannot establish any side-effects from the treatments, and conversely we also do not establish that the treatments do **not** cause side effects.

7 Privacy Considerations

As a final consideration for this paper, we also discuss the important privacy implications of the data. Our dataset does not contain directly identifying features such as subject name or post code. The data can therefore be considered anonymized, but it is notably not sufficient to guarantee that the individuals described by data could not be identified. The primary identifying features that would be expected to be publicly available are the subject's age and gender, as well as their income. Additionally, information regarding where the information was collected could be important to determining the risk to privacy. For instance, if this dataset were collected from the population of the United States, the risk would be relatively low of identifying any individual out of a population of more than 300 million, using only the age, gender, and income. However, if the dataset is drawn from a relatively smaller population such as the city of Oslo or even a less inhabited area, these features are comparatively more informative for an attacker attempting to identify individuals from the data. This is most likely if, for instance, the attacker knows the population the sample was drawn from, and could have access to other information which could be connected to the data.

One method that could make the data more secure, although less precise, would be to bucket the ages into brackets, rather than the precise value. This would reduce the distinctiveness of any one individual with respect to another when observing the age feature, and potentially establish a notion of k-anonymity in the data. (Sweeney, 2002) This is to say that, for any individual in the data, they will be indistinguishable from at least $k-1$ other individuals in the data (where we are treating all features as quasi-identifiers). As an example, if records are stored with the age feature represented as a bucket, then a 50 year old woman and a 59 year old woman could not be distinguished using this feature. This approach alone may not be likely to properly ensure k-anonymity due to the large number of other features, but the age feature may be comparatively more accessible as prior information to a possible attacker (as opposed to e.g. genes or symptoms). Finally, the limitations of k-anonymity itself may require alternative approaches (see Machanavajjhala et al. (2007) for discussion of so-called *homogeneity attacks* and *background knowledge attacks*, the latter of which relates particularly to the considerations described in the previous paragraph).

As a specific alternative, privacy could also be strongly secured by introducing a differential-privacy algorithm prior to access of the data by a researcher or analyst. In this case, we could define a measure ϵ of privacy which determines how secure the data will be. For instance, in the collection of statistics over the population, noise can be added to the data prior to the calculation of each statistic. This approach would also be advantageous because any processing of the data done on the output of ϵ-DP will also be ϵ-DP.

8 Conclusion

To conclude, we summarize the answers to the questions posed in Section 1 and illustrated in detail in Section 6. Our analysis is composed of results from the three methodologies. Taken together, we expect that these three mechanisms provide a robust indication of which features affect patient outcomes and which are unimportant, and provide useful conclusions to the questions posed in this paper. We first successfully establish the efficacy of these methodologies on synthetic data, demonstrating that they are able to illustrate the relations between features that we defined in the data generation.

When applied to the real data, our methodologies are able to provide the following conclusions. To our questions (1a.), we are able to successfully predict death or survival given genes and comorbidities in the data. The most explanatory features for predicting death in the $CovidPositive$ population were $Gene_{16}$, $Gene_{63}$, and $Gene_{27}$. Our second questions (1b. and 1c.) are answered with the observation that $Vaccine1$ is the most effective in preventing death, but also that all Vaccines show

an ability to decrease the probability of death. We also observe likely side-effects from these vaccines, specifically $Headache$ and $Fever$. The probability of these side-effects from all three vaccines were similar. To our final questions (2), we were unable to determine if any side-effects are caused by the treatments or what they may be. However, our methodologies do predict that $Treatment2$ is very likely to be superior to $Treatment1$ in preventing death.

References

Jay L. Devore and Kenneth N. Berk. *Modern Mathematical Statistics with Applications*, second edition.

Aurélien. Géron. 2017. *Hands-on Machine Learning with Scikit-Learn TensorFlow: Concepts, tools, and techniques to build inteligent systems*. O'Reilly Media, Inc.

Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es.

Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

John B. Walsh. 2011. *Knowing the Odds: An Introduction to Probability*, graduate studies in mathematics: v. 139 edition. American Mathematical Society.