Jai Mehrotra-Varma
12/14/22

Final Project Report

Research Question: Do hospitalized COVID-19 positive patients have a higher mortality rate than hospitalized COVID-19 negative patients as a result of their COVID-19 status?

Description: My research question is to determine whether hospitalized COVID+ patients have a higher mortality rate than hospitalized COVID- patients. I obtained a dataset with hospitalized COVID+ patients and a dataset with hospitalized COVID- patients. Age, gender, hypertension, and diabetes status were chosen to be included since prior studies have shown these factors to be some of the greatest contributors to mortality, and I wished to control for these variables to see if COVID status as a singular independent factor influenced mortality. To do this, I conducted both an odds ratio analysis and a propensity score matching analysis with mortality as the dependent variable and the only major unadjusted independent variable being COVID status.

Motivation: As a pre-medical undergraduate student, I am very interested in working in the clinical setting, and through my hospital work, I have been exposed to numerous patients with a variety of conditions. I am very interested in going into healthcare data science research in the future, and I have read several research studies involving hospital-related data. Since COVID-19 is such a pressing issue right now, I thought it would be interesting to design and implement a study focusing on COVID-19 patients, and use some of the advanced analysis techniques used in healthcare studies to analyze patient data.

Dependencies:
pandas==1.3.4
numpy==1.20.3
matplotlib==3.4.3
seaborn==0.11.2
statsmodels==0.12.2
plotly==5.11.0
sklearn==0.24.2

Installation:
pip install -r requirements.txt

Code: The code presented was written in a single Jupyter notebook file (entitled “Final Project Code”). It can be run by administering Cell → Run All. Please see the commented note on Jupyter Notebook regarding the csv files. The link to the repository is below:
https://github.com/biony1209/DSCI-510-Final-Project

Jai Mehrotra-Varma
12/14/22

Data: For my final project, I collected two sources of data - both Kaggle datasets. One dataset was composed of hospitalized COVID-19 positive patients, while the other dataset was composed of hospitalized COVID-19 negative patients. The COVID+ dataset had 566,602 unique patients, while the COVID- dataset had 1176 unique patients. Both datasets provide demographic and hospital-related information about each patient (ex: their age, gender, comorbidities, lab values, etc.). I filtered the data to just include 1) their age, 2) their gender, 3) their hypertension status, 4) their diabetes status, 5) their COVID status (which I eventually added manually to the pandas dataframe), and 6) their mortality outcome. The files were labeled as "covid.csv" and "noncovid.csv" and samples of these datasets were put into the data folder.

Both of my datasets were downloaded from Kaggle as CSV files. The original CSV files were much larger (COVID+ dataset had 566,602 unique patients, while the COVID- dataset had 1176 unique patients), so for the sample data, I created two new CSV files selecting 50000 patients and 1000 patients from the datasets (respectively) and saved them to the data folder. They give a good sample overview of what the data looks like (the full datasets can be downloaded from the links below). These datasets were different from my initial plan, since I had initially planned to do an environmental science related project regarding plastic pellet distribution and factory locations. The main reason I didn't initially decide to pursue a COVID-related project was because it was very difficult to find a patient dataset that included all of the variables that I wanted to adjust for in order to answer my research question. For instance, I found many COVID datasets that showed mortality, but I had no way of knowing if it was due to COVID or if it was a result of some other factor (such as the COVID+ patient cohort being much older than the COVID- patient cohort). Even the two datasets that I ended up using had inherent biases - for instance, the COVID- cohort that I used had an average age of 74 while the average age of the COVID+ cohort was 42.6, so it was essential for me to adjust for this in my analysis. Also, I could not find a comprehensive dataset that included both COVID+ and COVID- patients, so I had to create a new COVID status column in my pandas dataframe (assigning 1 as COVID+ for every COVID+ patient and 0 as COVID- for every COVID- patient). This issue meant that I could not adjust for the differences in location, as the patients data was taken from different hospitals, and it was difficult to adjust for that. The two datasets can be shown below:

Data Sources:

Dataset #1: Hospitalized COVID+ patients:
This dataset was filtered to just contain age, sex, hypertension, diabetes, mortality outcome, and COVID status. Some of the columns used multiple numbers, but all categorical variables were converted to a binary (0 and 1) system (which was required for some of the analysis models). For sex: 0 represents females and 1 represents males, for outcome: 0 represents living and 1 represents deceased, for hypertension: 0 represents not having hypertension and 1 represents having hypertension, for diabetes: 0 represents not having diabetes and 1 represents having

diabetes, and for COVID status: 0 represents not having COVID and 1 represents having COVID (in this case, all 1s since every patient is COVID+. This was created manually.)

	id	outcome	age	sex	hypertension	diabetes	covid_status
0	16169f	0	27	1	1	1	1
1	1009bf	0	24	1	1	1	1
2	167386	0	54	0	1	1	1
3	0b5948	0	30	1	1	1	1
4	0d01b5	1	60	0	0	0	1
...	...	...	...	...	...	...	...
566597	01ff60	0	58	1	1	0	1
566598	047cd1	0	48	0	1	1	1
566599	1beb81	0	49	0	1	1	1
566600	16fb02	0	43	0	1	1	1
566601	0021c9	0	65	1	0	0	1

Dataset #2: Hospitalized COVID- patients:
This dataset was also filtered to just contain age, sex, hypertension, diabetes, mortality outcome, and COVID status. Some of the columns used multiple numbers, but all categorical variables were converted to a binary (0 and 1) system (which was required for some of the analysis models). For sex: 0 represents females and 1 represents males, for outcome: 0 represents living and 1 represents deceased, for hypertension: 0 represents not having hypertension and 1 represents having hypertension, for diabetes: 0 represents not having diabetes and 1 represents having diabetes, and for COVID status: 0 represents not having COVID and 1 represents having COVID (in this case, all 0s since every patient is COVID-. This was created manually.)

Link to both datasets (can be downloaded from here or from Kaggle):

https://drive.google.com/drive/folders/1PUDcBkMP6aTHxYem3o38rn8gO4qVM1-a?usp=sharing

Jai Mehrotra-Varma
12/14/22

	id	outcome	age	sex	hypertension	diabetes	covid_status
0	125047	0	72	0	0	1	0
1	139812	0	75	1	0	0	0
2	109787	0	83	1	0	0	0
3	130587	0	43	1	0	0	0
4	138290	0	75	1	1	0	0
...	...	...	...	...	...	...	...
1172	171130	0	62	0	1	1	0
1173	101659	0	78	0	0	1	0
1174	162069	0	85	1	1	1	0
1175	120967	0	79	1	0	1	0
1176	107636	0	47	0	1	1	0

Dataset #1 Link:
https://www.kaggle.com/datasets/tanmoyx/covid19-patient-precondition-dataset?select=covid.csv
Dataset #2 Link:
https://www.kaggle.com/datasets/saurabhshahane/in-hospital-mortality-prediction

Analysis:

To conduct my analyses, it was essential for me to correct for the inherent differences in the datasets in some way. When doing basic statistical analysis on my COVID+ dataset and COVID- dataset, I discovered that there was a statistically significant difference in several of my variables (ex: age and sex), so I knew that it wouldn't be appropriate to just do simple comparisons (ex: chi-squared test, t-test, regression, etc.) without first accounting for these differences in some way. I needed to be sure that COVID status was a determining factor for differences in mortality, and it wasn't just because one group happened to have older individuals or some other major confounding difference. As a result, the two main analyses I decided to do were odds ratio analysis and propensity score matching. Odds ratios are defined as "measures of association between an exposure and an outcome" - in other words, if the odds ratio model gave us a value of 2.5 for mortality in COVID+ patients, we would be able to make the claim that COVID+ individuals were 2.5 times more likely to die than COVID- patients. Specifically, I performed what is known as an adjusted odds ratio which controls for other predictor variables in the model. From reading up on other research papers that have made use of the odds ratio, I discovered that most researchers utilize age and sex as the two major variables to control for, so I

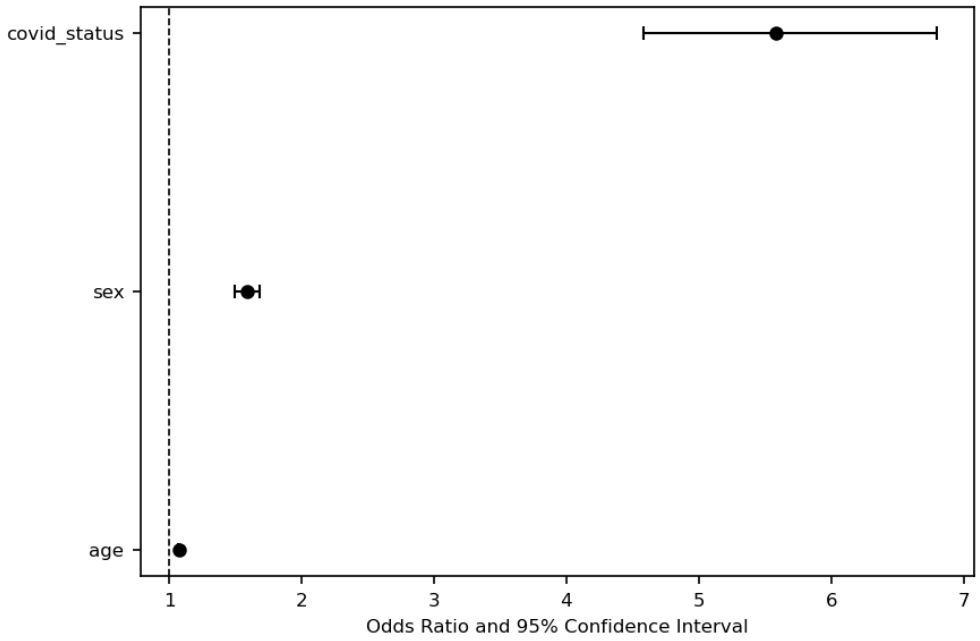
followed the same process and controlled for both age and sex while setting mortality as the dependent variable. My results were as follows:

	Odds Ratio	p-value	5%	95%	significant
covid_status	2.841336	8.996275e-32	2.386385	3.383022	True
sex	1.807317	0.000000e+00	1.765253	1.850383	True
age	1.071549	0.000000e+00	1.070798	1.072301	True

Logit Regression Results

Dep. Variable:	outcome	No. Observations:	50791
Model:	Logit	Df Residuals:	50787
Method:	MLE	Df Model:	3
Date:	Tue, 13 Dec 2022	Pseudo R-squ.:	0.1730
Time:	23:48:19	Log-Likelihood:	-15755.
converged:	True	LL-Null:	-19051.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.6295	0.127	-59.930	0.000	-7.879	-7.380
covid_status	1.7193	0.100	17.128	0.000	1.523	1.916
age	0.0705	0.001	71.191	0.000	0.069	0.072
sex	0.4615	0.030	15.232	0.000	0.402	0.521

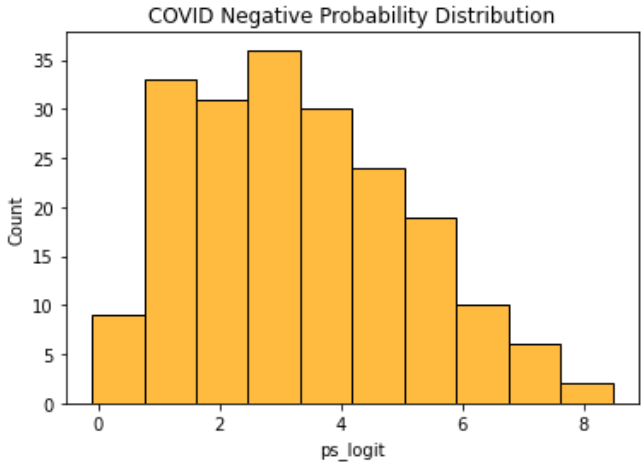


Ultimately, these figures show that age and sex were indeed contributing factors to the differences in mortality between the datasets (older individuals and males were more likely to die from COVID, shown by the p-values being less than 0.001), but even after these variables were

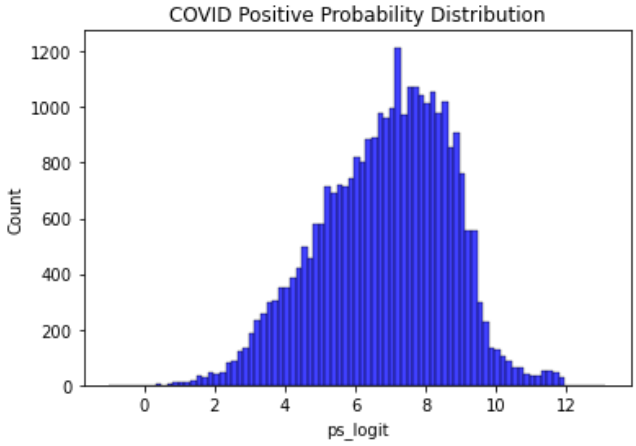
accounted for, COVID+ patients were still more likely to die than COVID- patients (odds ratio of 2.84). While the odds ratio analysis is undoubtedly an effective model, it works best when only a couple of the major variables are included (in this case - age and sex). I knew that diabetes and hypertension status could also significantly affect mortality rate, so to address the issue of these confounding variables, I wanted to expand my analysis to include them.

The other major analysis that I did is known as propensity score matching. This is another common technique used in COVID research that also addresses the influence of confounding variables and controls for them - ultimately creating a new "matched" cohort dataset that is not subject to the inherent biases of the confounding variables. For example, let's imagine that we have a COVID+ patient that is 85 years old, is a male, and has both diabetes and hypertension. If we compare that man to a COVID- patient who is 24 years old, is a female, and lacks diabetes and hypertension, and the 85-year-old man dies, the reason could have just as easily been due to age, sex, diabetes, or hypertension as opposed to his status as COVID+ (which is what I'm interested in). However, if I find another 85-year-old man who is 85 years old, is a male, and has both diabetes and hypertension **but** is COVID-, and only the COVID+ patient dies, we would be much more confident in accepting that COVID status may have played a significant role. In propensity score matching, a logistic regression is performed and propensity scores (defined as the probability of having/receiving a certain variable while adjusting for other covariates) are assigned to all of the individuals in both datasets based on the present variables. Since there are substantially more COVID+ than COVID- patients, I then matched 200 COVID- patients with 600 COVID+ patients (1:3 ratio) with the goal that each of the 3 COVID+ patients matched to each COVID- patient would be as similar as possible with regards to age, sex, diabetes status, and hypertension status (which is known by comparing propensity scores), with the only real difference being COVID status. This allowed me to create a new "matched" cohort that corrected for these other variables, similar to the adjusted odds ratio model. I was able to create some simple histogram visualizations to check if this technique worked by analyzing the distribution of the propensity scores for COVID- vs Unmatched COVID+ and COVID- vs Matched COVID+. The goal of creating a matched cohort is to pick individuals from one cohort that are similar in the distribution of every variable except the variable of interest. While the distributions for COVID- and Unmatched COVID+ were quite different, the distributions for COVID- and Matched COVID+ were quite similar, meaning that the matching was successful. The histogram distributions are shown below:

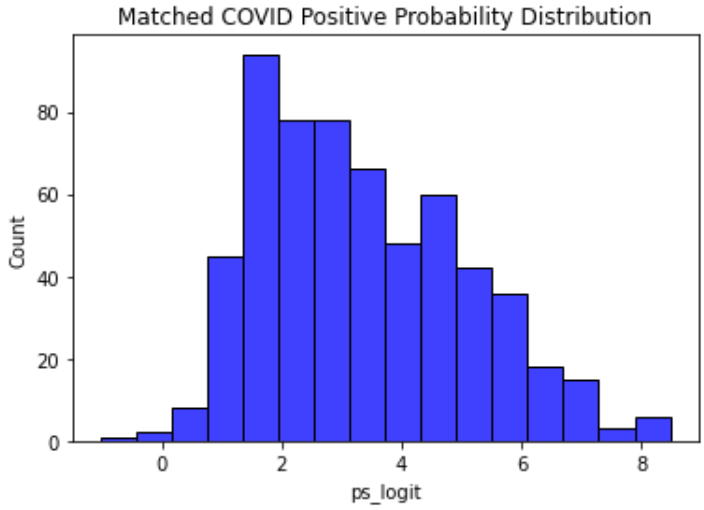
Jai Mehrotra-Varma
12/14/22



A.



B.



C.

As we can see from the figures, the shape of the distributions between A and B are different, while the shape of the distributions between A and C are very similar. This means that A and C must have a similar distribution of propensity scores (which is the variable plotted on the histograms). These figures thus verify that the propensity score matching was successful.

Jai Mehrotra-Varma
12/14/22

After creating a new matched COVID+ cohort, I was then able to compare the mortalities between COVID- and Unmatched COVID+ and COVID- and Matched COVID+. To do this, I performed chi-squared and t-test analyses to test for any statistically significant differences. The differences in age, diabetes status, hypertension status, and mortality for COVID- and Unmatched COVID+ were simply a reflection of the confounding variables discussed earlier, and they served as further evidence that it was important to correct for these differences:

	noncovid	covid	p-value	significant
Age - Mean	74.047619	42.601102	<0.001	True
Age - Standard Deviation	13.437241	16.651157	-	-
Male	618 (52.55%)	285830 (50.64%)	0.20	False
Diabetes	495 (42.09%)	493638 (87.46%)	<0.001	True
Hypertension	844 (71.77%)	472008 (83.63%)	<0.001	True
Mortality	159 (13.52%)	35888 (6.36%)	<0.001	True

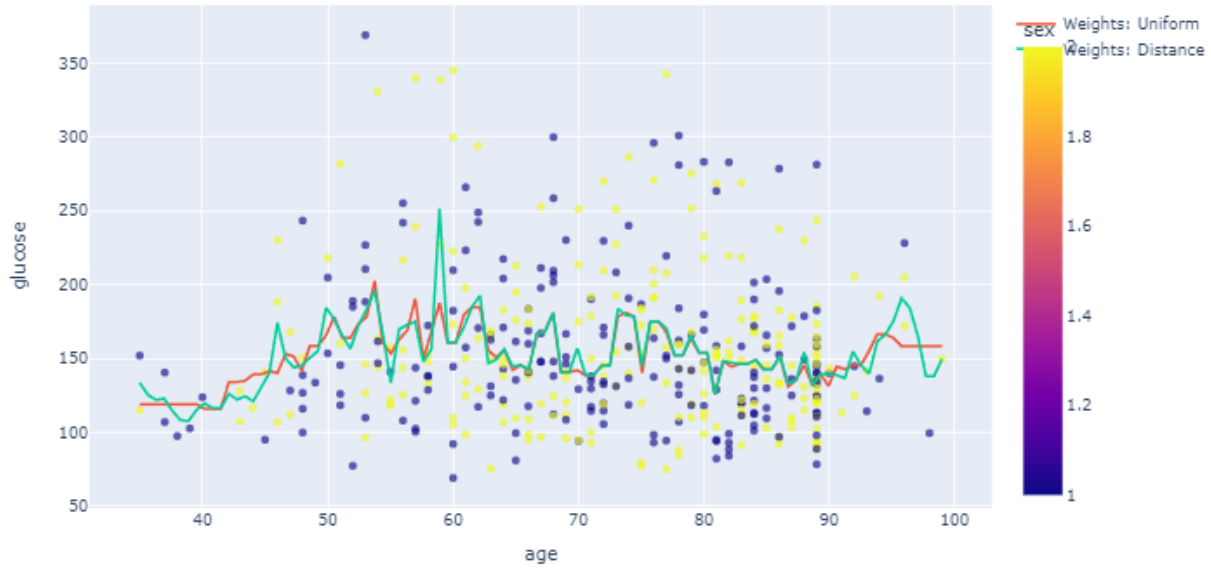
The important results were found when comparing COVID- and Matched COVID+, shown below:

	noncovid	matched_covid	p-value	significant
Age - Mean	74.047619	73.348333	0.30	False
Age - Standard Deviation	13.437241	13.39176	-	-
Male	618 (52.55%)	291.0 (48.5%)	0.12	False
Diabetes	495 (42.09%)	268.0 (44.67%)	0.32	False
Hypertension	844 (71.77%)	443.0 (73.83%)	0.39	False
Mortality	159 (13.52%)	206.0 (34.33%)	<0.001	True

As we expected, the propensity score matching ensured that there were not statistically significant differences in the age, sex, diabetes status, and hypertension status between the two cohorts. However, despite adjusting for all of these significant factors, there was still a statistically significant difference in mortality, with the matched cohort having a mortality rate of 34.3% and the COVID- cohort having a mortality rate of 13.5%. This significantly supports the idea that hospitalized COVID+ patients were more likely to die than hospitalized COVID-

patients due to differences in their COVID status (and not just because of other confounding variables). The results from the odds ratio and propensity score matching analysis were ultimately impactful because they both supported the idea that COVID status as a singular independent variable can increase mortality rates in hospitalized patients.

The biggest challenge I faced was with the runtime when performing propensity matching. Creating propensity scores for hundreds of thousands of patients ended up being unfeasible, but I realized that if I had at least 3 matched COVID+ patients for every COVID- patient, that would be sufficient to perform the matching analysis. As a result, I shortened the datasets to 30,000 COVID+ individuals and 200 COVID- individuals, which were large enough samples to conduct the analyses without being impractical. Another challenge that I faced was when I attempted to do correlation analysis between continuous variables (such as age and heart rate or age and blood glucose level, which were part of the original datasets that I ended up filtering out for my other analyses). I had expected to observe linear relationships between some of these variables, but I did not really observe any type of relationship. For advanced visualizations, I was particularly interested in looking at correlations in my data in a little more depth. However, instead of just doing a regular correlation plot, I decided to make use of the KNeighborsRegressor. This is an especially useful tool because when Linear Regression is typically performed, it strictly uses a linear model to predict the association between two variables. However, with this form of regression, we are able to estimate the association between the variables based on the proximity of the neighboring data points, which is more effective since in the real-world, data is typically not close to normal. I chose to do correlations between age and heart rate and age and blood glucose, since part of my earlier analysis was to control for these variables since I knew they were likely to affect other data fields (such as mortality). My visualizations were two interactive plots showing these two correlations and also dividing the data points into males and females to add another layer of depth. While I anticipated these would be valuable visualizations, I ended up not seeing significant relationships when doing this type of correlation, so I ended up not using this as a significant part of the analysis. Nevertheless, I included examples of these plots because I do think this analysis could be very useful for this type of study (although I would need to obtain the data for more continuous variables - not just categorical variables), and this is something I would have liked to expand on if given more time.



Future Work:

Ultimately, this project allowed me to implement some very practical analysis techniques and analyze datasets with COVID+ and COVID- patient data. If I was to expand the scope of this project, I think it would be valuable to adjust for the effects of even more confounding variables (not just the most significant ones that I looked at - age, sex, diabetes, and hypertension). For instance, there are other comorbidities that are sometimes adjusted for in COVID research studies - including chronic heart failure, chronic kidney disease, chronic obstructive pulmonary

disease, and cancer - that may have also influenced mortality in some way (which I could have adjusted for). Moreover, I would also like to repeat this study with larger and more comprehensive datasets. An issue that I had with my project is that obtaining datasets was difficult since most comprehensive patient datasets have restricted access (typically only granted to researchers working as part of hospital and/or medical school research division). If I was granted access, I could perform much more rigorous analyses (ex: more descriptive statistics on the cohorts).

Additionally, I would like to analyze multiple pairs of datasets from the same area, as my datasets looked at patients from different areas, and this may have impacted the results. Finally, I would like to expand the scope of the project from simply COVID+ vs COVID- (since this is a fairly basic question that has already been answered). There are a multitude of questions that have yet to be answered relating to how COVID can affect people with other conditions - as an example, recent research has shown that diabetic patients have much worse outcomes for COVID-19 than non-diabetic patients. I could repeat the analyses techniques in this project to look for deeper relationships within COVID+ and COVID- datasets, such as comparing the outcomes of COVID+ patients with lupus to COVID+ patients without lupus. With access to sufficient data, the possibilities are limitless to expand this project into something that substantially contributes to the field of COVID research.