Making Informed Decisions for Employee Health During COVID-19 Pandemic

Joshua Morgan 038
jmorgan63@gatech.edu

William Espinoza 039
wmespi@gatech.edu

April 21 2020

ISYE 6414: Regression

Abstract

SARS-COV-2, the virus that causes COVID-19, has been the most prevalent and deadly virus since the Spanish flu. While there are many reports detailing the epidemiology, these are often targeted at medical audiences and policy makers not managers. In this report, a combined regression analysis and cost benefit analysis framework is presented to help managers make informed decisions during the pandemic. From the regression analysis, a significant model was found with a F1 score of 0.71 indicating adequate performance. Age and chronic conditions strongly predicted death, and sex did not. Current recommendations highlight the need to mitigate risk for elderly individuals, and this analysis suggest that middle age individuals with chronic conditions should also mitigate their risk. Using the regression analysis a framework for the calculation of cost of infection was created which can easily translate to a cost analysis for managers. The cost calculation provided three main managerial insights. First managers should engage their employees about chronic conditions and suggest risk mitigation practices such as work from home. Second, managers should take into account the cost of staffing decisions because a high contact role have higher rates of infection at risk individuals should be removed from those roles. A cost benefit analysis tool was created to encourage managers to investigate alternatives in the face of risk. Finally, managers should use these conditions to develop talent as younger individuals face lower risk and could learn new skills while protecting their coworkers.

Contents

1 Introduction

1.1 Problem Description

As of April 19, 2020 SARS-CoV-2, the virus that causes COVID-19, has infected at least 2.4 million people world wide and has resulted 160,000 deaths [9]. SARS-CoV-2 is the most widespread and severe pandemic in the world since the Spanish flu pandemic of 1918. Original forecasts from experts forecasts around 2 million deaths in the United States, and in response to this this existential threat, policy makers worldwide have responded with social distancing policies to mitigate the exponential spread of the disease. [10]. Fortunately, the first wave of infections have peaked with the IHME model predicting the United States is four days past the peak [6]. While social distancing has flattened the curve and saved thousands of lives, the economic cost have been staggering. Goldman Sachs recently forecast a 3.8 percent decrease in Gross Domestic Product or a loss of 814 billion dollars [11]. In the face of such staggering economic loss and death, policy makers and managers are faced with tough decisions on how to protect public health and the livelihood of all Americans.

1.2 Motivation

Currently, managers can access local disease prevalence, government guidelines, and their own operational knowledge, but they often exist in terms of probabilities, statements, or managerial intuition. In *Thinking Fast and Slow*, Kahneman condense describes humans as irrational decision makers when confronted with abstract probabilities tending to think in absolutes such as X will never happen or X will happen [12]. Furthermore, individuals tend to overweight the consequences of financial losses than financial gains let alone health outcomes [19]. Since people are often irrational decision makers, government guidelines and raw public health data can be misunderstood or misconstrued. Instead of making a detailed analysis of multiple options, people often use heuristics or quick fast rules to make decisions. While public health guidelines are straightforward, they often do not encourage rational decision, so all material for managers should be synthesized to prevent heuristics from dominating decision making. Fortunately, managers with formal business training have been schooled in cost benefit analysis, so by framing different decisions in terms of costs could prevent the predominance of heuristics. While cost benefit analysis are effective decision making procedures, they often can be difficult to create or raise ethical questions, but in the fog of the pandemic a easy to use cost benefit analysis system could provide the public with a clear way to make decisions. To synthesize the abundant information about COVID-19, a simple cost benefit calculator combining local public health statistics, government guidelines, and employee information can be used to quantify and to monetize the impact of different decisions on employees.

1.3 Challenges

At the current stage of the pandemic, data sources are often incomplete and contain few features useful for predicting outcomes. Studies are often commissioned by the government for policy makers and contain few managerial insights and clinical data. Managers have access to simple questions about age, sex, and the presence of chronic conditions not vital signs. Under these constraints, a significant analysis could be difficult to perform or contain inaccuracies.

1.4 Approach

While informing managerial decisions such as staffing is difficult, we first seek to quantify the costs of an infection in terms easy to understand for business professionals. First, our data acquisition and preprocessing strategy is described, and the, a regression is used to predict the outcome from easily accessible data. Next the probability of a given outcomes used from our regression analysis to predict the outcome for a given individual. With these outcomes and probabilities, a methodology for quantifying the effects of a COVID-19 infection will be described. Finally, general trends and findings will be reported from our analysis.

2 Data Description

In order to build a model that would predict the outcome of a COVID-19 patient, we had to acquire a data set, specifically with information on an individual case basis. While many of the available data sets are more oriented for time series modeling and only have aggregated case counts, the Hubei Province in China had a de-identified COVID-19 patient data set [8]. This data set contained a information for 261,558 cases with a wide range of variables. However, because there was a large proportion of cases with incomplete information, extensive cleaning and dimension reduction was required. Appendix A.1.1 contains an in depth explanation of these cleaning and filtering mechanisms. After cleaning the data, the data set was reduced to 7 variables (age, sex, latitude, longitude, country, chronic disease status, outcome) and 413 cases. Age, latitude, and longitude were all continuous variables. Sex, country, chronic disease status, and outcome were all categorical variables. Because chronic disease status was a valuable variable with many missing values, it was encoded as a categorical variable with 3 possible values, positive, negative, or unknown. The rationale for this decision can be found in Appendix A.1.2. The code used to execute this can be found in Appendix A.1.3. The outcome variable was limited to only patients who were not still in a COVID-19 episode. In this final data set of 413 cases, 120 did not survive their COVID-19 episode and 289 did.

3 Proposed Methodology

3.1 Regression Modeling

Because this problem requires a binary classification on the outcomes of surviving or dying, a logistic regression model was chosen. This model included all variables listed in the Data Description section with age, sex, and chronic disease status serving as the predictor variables, latitude, longitude, and country serving as control variables, and outcome serving as the dependent variable. The different features were first checked for significance by implementing the model using the python module statsmodels [18]. Following the feature and model analysis, a model was built in Python sci-kit learn using the logistic regression module [16]. To validate the model, a initial 75-25 train test split algorithm was used to insure that the model performed well. To determine if the model was robust, a 10 fold cross validation procedure which shows the models consistency across different sets of data. We decided to use the F1 score as our metric for this classification problem since it provides a easy summary of the accuracy. Appendix B.1 provides a more detailed description of regression modeling decisions and choices. Following the modeling, odds ratios were calculated for each variable, and the coefficients were used for the cost analysis.

3.2 Cost Analysis Framework

Our regression model predicts a outcome and also yields a probability of occurrence. These probabilities are quite small and their respective outcomes can have quite large consequences. To help managers understand these terms, the probability of death is multiplied with a cost. To calculate this cost or the cost of death, two models are used either a quality of life year (QALY) or a constant value (CV) [17]. For QALY, economic and health literature usually assign a value between \$50,000 and \$150,000 per year, and for a constant value around \$9 million [15] [13]. Appendix B.2 provides a detailed description of these two calculations.

4 Analysis and Results

4.1 Regression

Our initial train test split approach indicated that our model was adequate due to a good F1 score of 0.73 indicating that the model had a decent accuracy considering our unbalanced data. The 10 fold cross validation F1 score of .7171 indicates that the model has adequate predictive power, but it performed much stronger with specificity than sensitivity (Appendix D. Figure 3). Hypothesis tests were conducted for predictor variables to determine which were statistically significant (Appendix D. Figure 1). For a detailed description of the procedure of hypothesis testing results see Appendix D.1. From the hypothesis, testing age and chronic disease were the most important indicators each being statistically significant. The coefficient of age was .061 (OR=1.06), which indicated an increased risk as age increases. The significance of age has been documented in the literature and reiterated by the public health guidelines, and our model adds validity to this as well. The coefficient of the positive chronic disease status was 1.571 (OR = 4.8156) which indicated a relative risk of 6.87 compared to the reference category of no knowledge of a case's chronic disease history. The negative chronic disease variable was not significant, but overall chronic disease remains a significant variable. The strong association between chronic disease and risk of death has been documented in the literature; however, public health guidelines have focused more on age than chronic disease. Our model indicates that chronic diseases should be stratified as data becomes available and could be added to criteria for continued social distancing. Our model has indicated that sex is not a predictive factor for death, and this increases evidence for findings in the medical literature that there are no sex predispositions [5]. It could also highlight the strong binding between the COVID-19 spike proteins and angiotensin-converting enzyme 2 (ACE2) found in the medical literature [20].

4.2 Cost Analysis Framework

Our cost analysis framework showed the benefit of collapsing probabilities into simple costs. Our results for the QALY calculation with probability of death showed that guidelines suggesting those with chronic conditions found that costs increase as age increases indicating that younger individuals even with chronic condition shave less to lose with an infection. The total cost peaks at 60 years old with a cost of infection of $ 2 million and tapers off for older individuals (Appendix D, Figure 4). The high cost of infection for those over the age of 50 with chronic conditions is greater than $ 1.5 million indicating that these individuals face major costs still with a infection and likely should maintain shelter in place to mitigate the risk of infection. Individuals such as this in a high contact high volume roles such as a cashier at a fast food restaurant should avoid such situations. Moreover, individuals in the presence of asymptomatic carriers such as the young should also mitigate risk by reducing contact.

The constant value of life calculation found similar results without the later age tapering(Appendix D, Figure 6). Due to the higher value of life at $ 9 million the costs are greater; however, the chart should be interpreted more for general trends than actual sums. The risk faced by infection trends remain similar with high costs for individuals with preexisting conditions despite being middle aged. These high costs give weight to policy maker's decisions to encourage these individuals to stay home.

Our two methods of cost analysis lead to similar conclusions and suggest that managers should find ways to mitigate infection if possible. Our cost calculator helps managers calculate the cost of working by including a chance of infection (Appendix D, Figure 5). This helps managers weight in the quantity of contacts faced in some roles. Our quantification strategy leads to similar guidance reached by policy makers and public health officials. while the validity and ethical implications of such calculations are questionable, they provide a easy way to understand the consequences of decisions.

5 Conclusion

The findings from the logistic regression results have strong implications. First, the model's ability to adequately differentiate between cases that are likely to survive and those that are not suggests that in its current form it could be used to assess a prospective cases risk of dying if they were to contract the disease. Additionally, the predictor variables' coefficients allow immediate interpretation for which kinds of groups are most at risk. The findings about age support the current literature's assessment of higher age groups being at higher risk. A positive chronic disease status has not been as widely discussed as an at risk group, but the findings from this study suggest that this is an important consideration. Second, even with the limited data available, the findings from this model suggest that an even more robust model could be created as more data comes to light. A more in depth exploration of the chronic disease variable could provide insight into what kinds of chronic disease groups have the highest risk of dying should they contract the disease. Furthermore, the outcome variable could be expanded beyond just a binary classifier which would allow for the exploration of severity of outcomes as opposed to just survival. Public health officials should start to explore different outcomes because even though the virus might leave the body physiological damage from Acute Respiratory Distress syndrome which can require months of physical therapy. Our regression faced severe limitations because of noisy and limited data. The Hubei data set had many more cases available, but unfortunately a very small proportion of them were usable. Going forward this study can be used to inform which variable are important to track for COVID-19 patients (age, chronic disease status). This kind of information can be used to minimize the risk of the returning workforce by focusing isolation on the most at-risk groups.

Our cost analysis found three main managerial insights. First, managers should ask individuals whether or not they have a chronic disease such as chronic obstructive pulmonary disorder, diabetes, hypertension, asthma etc. These different risk factors are important for understanding the potential costs of disease, and they are often not shared with employers. Employees and employers should realize that under these extreme circumstances information is a weapon to fight risk and should share and protect this information. Managers should investigate risk mitigation strategies that balance business continuity such as work from home. Second, managers should understand the cost of their staffing decisions. Since age and chronic conditions are both strong predictors of death on infection, managers should realize the potential costs of sending individuals with these risk factors on "essential" work related travel or in "essential" roles. Managers should explore alternatives for these individuals to protect business continuity and use a simple cost benefit analysis to try and comprehend the impact of their decisions. The cost benefit calculator presented provides a easy tool for managers to understand the cost while they can easily calculate the costs. Finally, managers should use the pandemic as a opportunity to develop human capital in younger employees while protecting their high risk employees. Moreover, for future pandemics communication to managers should be investigated since managers often have direct control over the amount of contact different individuals have during the workday.

5.1 Class Feedback

This semester has been a great learning experience for the both of us. This project has taught us how to make a useful project and tackle issues such as missing data in a regression analysis. Our literature review on missing data introduced both of us to various concepts and ideas about tackling this issue. This project, inadvertently, taught both us valuable lessons about working remotely on different schedules. We had both taken the undergraduate version of this and were worried that material would be rehash and slow; however, the class has taught both us a much deeper understanding of regression methodologies. After completing this course, we both feel prepared for our future jobs in industry.

A Appendix: Data

A.1 Preprocessing

A.1.1 Filtering

Initially, all cases with any missing variable were filtered out. This left only a few cases available to build the model, so it was determined that the dimensions of the data set would need to be reduced. The remaining variables that would leave an adequate sample size were age, sex, latitude, longitude, country, chronic disease status, and outcome. Except for age, latitude, and longitude, all of these were encoded as categorical variables with the reference categories being the variable level with largest proportion of cases. After cleaning out all null values for the remaining variables, the data set now contained 495 cases. However, additional filtering was required on the outcome variable. This variable had 10 possibilities which were then grouped into 3 categories: dead (death, died, dead, deceased), episode in progress (under treatment, stable), and survived (discharged, released from quarantine, recovered, alive). Because the goal of the model is to predict the outcome at the end of a COVID-19 episode, all of the in progress cases were removed from the data set. This left 413 cases for the final data set. For the implementation of the filtering processes see Appendix C Listing 1.

A.1.2 Missingness

Unfortunately, at this stage of the pandemic data is often missing for individuals due to the rapid progression of the disease and limited public health resources. Records will often contain the age, sex, and outcome, but not include the presence of a chronic condition or include a unknown value. In health informatics literature, missingness is usually handled through either a imputation strategy such as multivariate imputation by chained equations (MICE) [4], K-nearest neighbors imputation (KNN) [3] or creating a new value for missingness [1] [7] [14]. For the pandemic case scenario with few supporting features imputation strategies such as MICE and KNN have very few features to provide a accurate estimate, and with limited public health resources officials might only record the easy to access information. Consequently, we assume that missing data is likely caries information about an individual, so we will encode a missing entry into a categorical variable.

B Appendix Methodology

B.1 Regression

The logistic regression package in sci-kit learn uses a fisher information score to solve the model and implements a L-1 regularization procedure as well. A 75-25 train test split for the algorithm was initially chosen because of simplicity and ease of interpret-ability. Following a adequate data set and modeling framework, a 10 fold cross validation was used to test for the robustness. To see the implementation of this see Appendix C Listing 2.

B.2 Cost Benefit Analysis

B.2.1 Quality of Life Year Calculation

To calculate the cost under a quality of life year calculation, first the number of life years lost must be calculated. To compute the number of life years lost, the United States life tables which describe the number of years a individual is expected to live at a certain age [2].

We will now formalize this in relation to COVID-19. We let A be the age of an infected individual. We let T be the time of death. We let H be the health status where $H = 1$ indicates a chronic and $H = 0$ is the lack of a chronic condition. We let C be the cost, and we let V be the value of a life year.

$$L = E[T|A] - A \tag{1}$$

$$E[C] = V(P(T = A|H)(E[T|A] - A)) = V(P(T = A|H)L) \tag{2}$$

Equation (1) describes the calculation of the life years, and we obtain $E[T|A]$ from the life tables. Equation (2) describes the expected cost given an infection. These two equations provide a efficient simple way to monetize the cost of infection.

B.2.2 Constant Value of Life Calculation

The calculation of a constant value of life is presented in equation 3, and provides a much simpler calculation for cost of infection.

$$E[C] = V(P(T = A|H)) \tag{3}$$

C Appendix Code

```python
import pandas as pd
import numpy as np

def main(path):
    new_cols = ['age','sex','chronic_disease_binary','country','latitude','longitude','outcome']
    data = pd.read_csv(path)
    data = preprocess2(data,new_cols)
    data.to_csv('cleaned_covid.csv')
    return data

def preprocess(data,new_cols):

    ## restrict starting columns
    df = data.copy()#[new_cols]
    df.chronic_disease_binary = df.chronic_disease_binary.fillna('')
    df['chronic_disease_binary'] = df['chronic_disease_binary'].apply(lambda x: adjust_chronic(x))
    df = df[new_cols].dropna()

    ## clean age column
    counts = []
    for i,age in enumerate(df['age']):
        try:
            int(age)
            counts.append(age)
        except ValueError:
            continue

    df = df[(df['age'].isin(counts))]
    df['age'] = df['age'].astype(int)


    ## clean outcome column
    df['outcome'] = df['outcome'].apply(lambda x: adjust_outcome(x))

    ## encode categorical variables
    cat = ['chronic_disease_binary','sex','country','outcome']
    df = pd.get_dummies(df,columns=cat)

    ## drop reference cases
    ## sex reference case: male, 291 cases
    ## country refernce case: China, 136 cases
    ## chronic disease reference case: unknown, 321 cases
```

```
    ## outcome reference case: Survive episode, 289 cases
    rcs = ['sex_male','country_China']
    df = df.drop('sex_male',1)
    df = df.drop('country_China',1)
    df = df.drop('outcome_S',1)
    df = df.drop('chronic_disease_binary_U',1)

    ## clean out in progress cases
    df = df[df['outcome_IP'] != 1]
    df = df.drop('outcome_IP',1)

    return df

def adjust_outcome(outcome):
    dead = ['death','died','Dead','Deceased','Died']
    IP = ['stable','Under treatment','Stable']
    lived = ['discharged','discharge','released from quarantine',
             'Discharged','recovered','Alive','Recovered']

    if outcome in dead:
        return 'D'
    if outcome in IP:
        return 'IP'
    if outcome in lived:
        return 'S'

def adjust_chronic(chronic):

    if chronic == '':
        return 'U'
    if chronic == 0:
        return 'N'
    else:
        return 'Y'

main('covid.csv')
```

Listing 1: Filtering Code

```
from sklearn.model_selection import KFold
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
import matplotlib.pyplot as plt
y = df.outcome_D
X = df.drop('outcome_D',axis=1)
kf = KFold(n_splits=10)
conf_matrix = []
f1 = []
auc = []
for train, test in kf.split(df):
    mod = LogisticRegression(solver='liblinear')
    mod.fit(X.loc[train,:],y.loc[train])
    pred = mod.predict(X.loc[test,:])
    conf_matrix.append(metrics.confusion_matrix(y.loc[test],pred))
    f1.append(metrics.f1_score(y.loc[test],pred))
print(sum(conf_matrix)/10)
```

```
print(sum(f1)/10)
```

Listing 2: Logistic Regression Implementation

C.1 Code link

In this section, we present a few important pieces of our software implementation for this project. For a complete, view of our code visit our https://github.com/jemorgan1000/Regression_Final_Project.

D Appendix Results

D.1 Regression

Hypothesis tests were conducted for predictor variables to determine which were statistically significant (Appendix D. Figure 1). For each of these tests, the null hypothesis states that the coefficient for the variable is equal to zero while the alternative hypothesis states that the coefficient has a non-zero value (Appendix D. Figure 1). For these tests, a standard significance level of .05 was used. The p-value of age and the positive chronic disease status categorical variable was 0 through three decimal points, indicating high significance for predicting the outcome of a case. While the negative chronic disease status was not significant with a p-value of .999, the underling variable of chronic disease status remains a significant variable to the model.Sex had a p-value of .155, indicating that it is not a significant predictor.

D.2 Figures

Generalized Linear Model Regression Results

Dep. Variable:	outcome_D	No. Observations:	413
Model:	GLM	Df Residuals:	385
Model Family:	Binomial	Df Model:	27
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-80.861
Date:	Tue, 21 Apr 2020	Deviance:	161.72
Time:	12:37:51	Pearson chi2:	324.
No. Iterations:	24		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
age	0.0842	0.015	5.795	0.000	0.056	0.113
latitude	-0.1572	0.063	-2.493	0.013	-0.281	-0.034
longitude	0.0075	0.017	0.436	0.663	-0.026	0.041
chronic_disease_binary_N	-24.8671	2.16e+05	-0.000	1.000	-4.23e+05	4.23e+05
chronic_disease_binary_Y	1.9953	0.528	3.779	0.000	0.961	3.030
sex_female	-0.8735	0.448	-1.952	0.051	-1.751	0.004
country_Algeria	1.5527	2.397	0.648	0.517	-3.145	6.250
country_Australia	-13.0870	4.725	-2.769	0.006	-22.349	-3.825
country_Brazil	-33.3102	1.01e+05	-0.000	1.000	-1.98e+05	1.97e+05
country_France	3.5190	3.265	1.078	0.281	-2.880	9.918
country_Gambia	19.9129	2.16e+05	9.22e-05	1.000	-4.23e+05	4.23e+05
country_Germany	-22.8831	2.16e+05	-0.000	1.000	-4.23e+05	4.23e+05
country_Ghana	2.697e-14	1.38e-10	0.000	1.000	-2.7e-10	2.7e-10
country_Guyana	21.5614	2.16e+05	9.98e-05	1.000	-4.23e+05	4.23e+05
country_Italy	3.7123	2.969	1.250	0.211	-2.108	9.532
country_Japan	-24.0155	1.07e+05	-0.000	1.000	-2.1e+05	2.1e+05
country_Malaysia	-29.8823	7.49e+04	-0.000	1.000	-1.47e+05	1.47e+05
country_Nepal	-24.5473	1.53e+05	-0.000	1.000	-2.99e+05	2.99e+05
country_Niger	-5.1193	1.481	-3.456	0.001	-8.022	-2.216
country_Philippines	-4.4098	1.304	-3.382	0.001	-6.966	-1.854
country_Romania	-20.9343	8.13e+04	-0.000	1.000	-1.59e+05	1.59e+05
country_San Marino	24.9649	2.16e+05	0.000	1.000	-4.23e+05	4.23e+05
country_Singapore	-29.9411	2e+04	-0.001	0.999	-3.92e+04	3.92e+04
country_South Korea	-24.1229	5.81e+04	-0.000	1.000	-1.14e+05	1.14e+05
country_Switzerland	-21.8192	1.47e+05	-0.000	1.000	-2.88e+05	2.88e+05
country_Thailand	-28.4548	1.53e+05	-0.000	1.000	-2.99e+05	2.99e+05
country_United States	26.6444	1.53e+05	0.000	1.000	-2.99e+05	2.99e+05
country_Vietnam	-29.9712	3.04e+04	-0.001	0.999	-5.96e+04	5.95e+04
country_Zimbabwe	19.9999	2.16e+05	9.26e-05	1.000	-4.23e+05	4.23e+05

Figure 1: Statistical Results from logistic regression

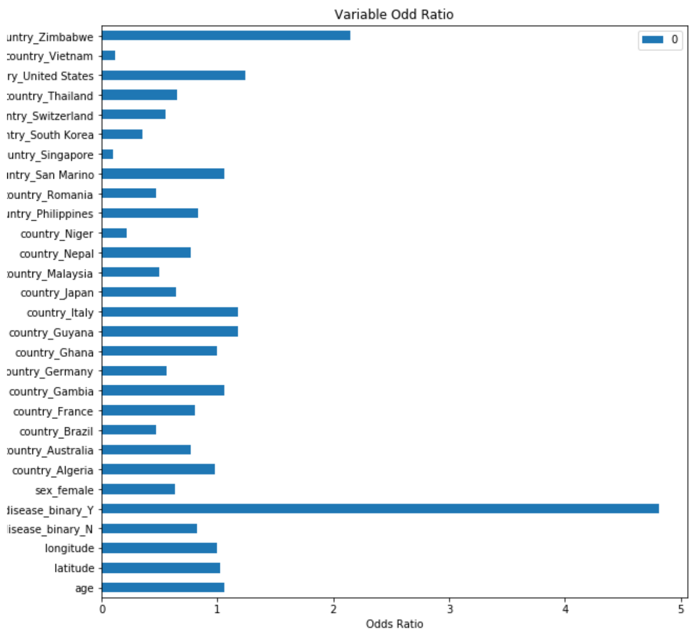

Figure 2: Odds Ratios for each variable in the regression.

	True Positive	True Negative
Predict Positive	25.9	3.4
Predicted Negative	3.4	8.6

Figure 3: Confusion Matrix from 10 fold cross validation

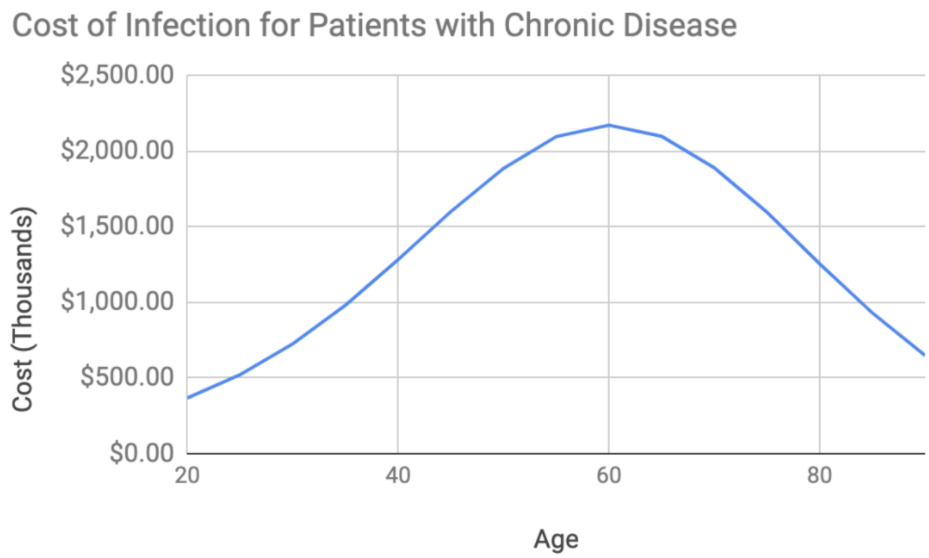

Figure 4: Cost Benefit Analysis with Quality of Life Year Calculation

Cost Benefit Analysis of Working Calculator	
Number of Local Cases	1,000
Area Population	500,000
Number of Customers Per day	1000
Expected Contacts	2
Prob Infection	0.1
Employee Age rounded down to 5	25
Employee Chronic Conditions	1
Employee Sex	1
Prob of Death with infection	0.0001
Prob of Death from working	0.00002
Life Years Remaining	54.7
Cost of Death	5470000
QALY	100,000
Cost of Working	$109.40

Figure 5: Cost Benefit Analysis calculator

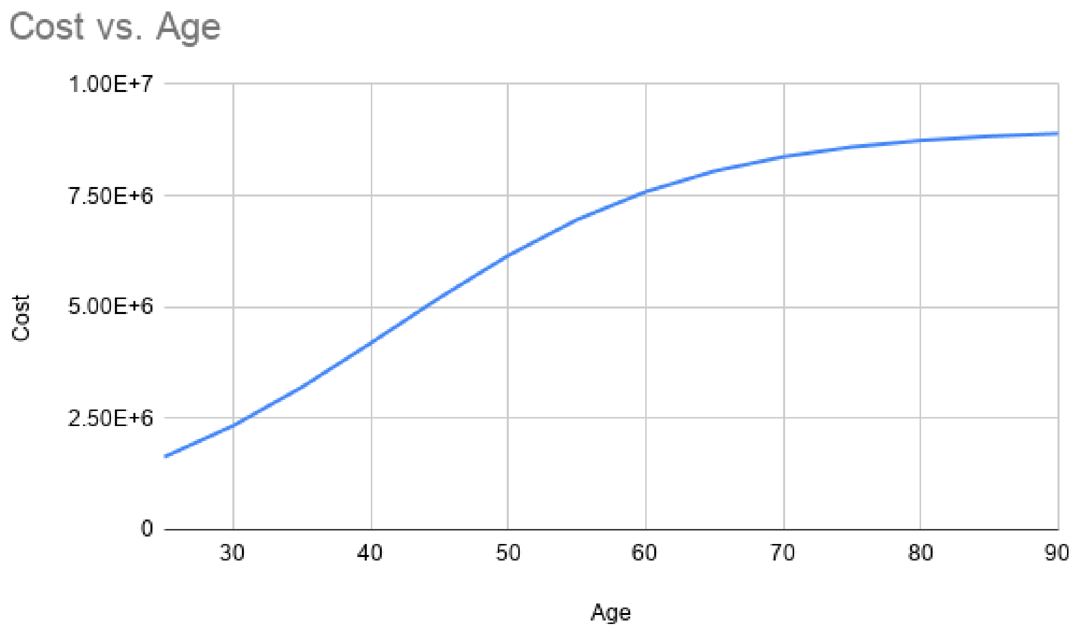

Figure 6: Cost Benefit Analysis with constant value of life

References

[1] Paul S Albert and Dean A Follmann. "A random effects transition model for longitudinal binary data with informative missingness". In: *Statistica Neerlandica* 57.1 (2003), pp. 100–111.

[2] Elizabeth Arias. "United States life tables, 2012". In: (2016).

[3] Lorenzo Beretta and Alessandro Santaniello. "Nearest neighbor imputation algorithms: a critical evaluation". In: *BMC medical informatics and decision making* 16.3 (2016), p. 74.

[4] S van Buuren and Karin Groothuis-Oudshoorn. "mice: Multivariate imputation by chained equations in R". In: *Journal of statistical software* (2010), pp. 1–68.

[5] Hua Cai. "Sex difference and smoking predisposition in patients with COVID-19". In: *The Lancet Respiratory Medicine* 8.4 (2020), e20.

[6] IHME COVID, Christopher JL Murray, et al. "Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months". In: *medRxiv* (2020).

[7] Desmond Curran et al. "Identifying the types of missingness in quality of life data from clinical trials". In: *Statistics in medicine* 17.5-7 (1998), pp. 739–756.

[8] nCoV-2019 Data Working Group. *Epidemiological Data from the nCoV-2019 Outbreak: Early Descriptions from Publicly Available Data*. Accessed on yyyy-mm-dd from http://virological.org/t/epidemiological-data-from-the-ncov-2019-outbreak-early-descriptions-from-publicly-available-data/337. 2020.

[9] Ensheng Dong, Hongru Du, and Lauren Gardner. "An interactive web-based dashboard to track COVID-19 in real time". In: *The Lancet infectious diseases* (2020).

[10] Neil Ferguson et al. "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand". In: (2020).

[11] Jan Hatzius, Alec Phillips, and David Mericle. *Insights - US Daily: A Sudden Stop for the US Economy*. Mar. 2020. URL: https://www.goldmansachs.com/insights/pages/us-daily-20-march-2020.html.

[12] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[13] Chris P Lee, Glenn M Chertow, and Stefanos A Zenios. "An empiric estimate of the value of life: Updating the Renal Dialysis Cost-Effectiveness Standard". In: *Value in Health* 12.1 (2009), pp. 80–87.

[14] Jau-Huei Lin and Peter J Haug. "Exploiting missing clinical data in Bayesian network modeling for predicting medical problems". In: *Journal of biomedical informatics* 41.1 (2008), pp. 1–14.

[15] Molly J Moran and Carlos Monje. *Guidance on Treatment of the Economic Value of a Statistical Life in U.S. Department of Transportation Analyses*. Department of Transportation, Aug. 2016. URL: https://www.transportation.gov/sites/dot.gov/files/docs/2016%20Revised%20Value%20of%20a%20Statistical%20Life%20Guidance.pdf.

[16] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[17] Luis Prieto and José A Sacristán. "Problems and solutions in calculating quality-adjusted life years (QALYs)". In: *Health and quality of life outcomes* 1.1 (2003), p. 80.

[18] Skipper Seabold and Josef Perktold. "Statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[19] Amos Tversky and Daniel Kahneman. "Advances in prospect theory: Cumulative representation of uncertainty". In: *Journal of Risk and uncertainty* 5.4 (1992), pp. 297–323.

[20] Muthiah Vaduganathan et al. "Renin–angiotensin–aldosterone system inhibitors in patients with Covid-19". In: *New England Journal of Medicine* (2020).