December 2020

Estimating Covid-19 Survival Rate and Inferring Case Severity with Respect to Milwaukee County Policy Change Using Logistic Regression

Geoff M. Chappelle
University of Wisconsin-Milwaukee

ESTIMATING COVID-19 SURVIVAL RATE AND INFERRING CASE SEVERITY WITH RESPECT TO MILWAUKEE COUNTY POLICY CHANGE USING LOGISTIC REGRESSION

by

Geoff Chappelle

A Thesis Submitted in
Partial Fulfillment of the
Requirements for the Degree of

Master of Science
in Biostatistics

at
The University of Wisconsin-Milwaukee
December 2020

ABSTRACT

ESTIMATING COVID-19 SURVIVAL RATE AND INFERRING CASE SEVERITY WITH RESPECT TO MILWAUKEE COUNTY POLICY CHANGE USING LOGISTIC REGRESSION

by

Geoff Chappelle

The University of Wisconsin-Milwaukee, 2020
Under the Supervision of Professor Shengtong Han, Ph.D.

Coronavirus disease 2019 (COVID-19) is a global issue, and it is affecting 170 countries in very different ways. In the United States, a lot of efforts have been made nationally and by individual states to curb the spread and severity of COVID-19. Policy changes and recommendations have been met with variable success across the country. There is a wealth of information on where COVID-19 infection and death are prevalent, and there are several articles discussing disparities in those outcomes among different populations. However, those findings are not necessarily tied to a policy change or, in the weeks that follow a change, a description of the corresponding change in COVID-19 prevalence and severity, if any. In this thesis, we will use univariate logistic regression and the cumulative logit model to identify the population in Milwaukee County most at-risk for death from COVID-19 with respect to age, race, and sex, using confirmed COVID-19 case and death data from the Wisconsin Department of Health Services. We will then break the data apart into time intervals of approximately two months to see if these risks were more or less severe as a function of policy changes made regarding social distancing, requiring a mask, and limiting non-essential work interactions.

To my wife and our first daughter, expected graduation to the world April 2021.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

The SARS-CoV-2 (coronavirus, COVID-19) outbreak has become a global pandemic. The first cases of ‘viral pneumonia’ were reported by the Wuhan Municipal Health Commission in Wuhan, China, in late December 2019.The first Disease Outbreak News report was issued by the World Health Organization (WHO) on January 5, 2020 [1]. In the next few days, Chinese authorities determined that the outbreak was caused by a novel coronavirus. The WHO began evaluating the infectivity and reach of the disease and, weeks into their investigations, declared the outbreak a public health emergency of international concern (PHEIC), which is the highest level of alarm.

It was not long before the coronavirus reached other countries. A patient in the Washington, United States, was publicly confirmed to be positive for COVID-19 at the end of February. The first coronavirus case in Wisconsin was confirmed on February 5th [2]. The months following these first confirmed cases showed a steady rise in coronavirus across the United States and the entire world. Some places were affected worse than others with respect to both infection by and mortality from COVID-19. While the CDC was instituting national-level guidelines, such as social distancing and avoiding large gatherings, similar policies were underway at the state level.

Timeline of Events Related to COVID-19 Progress in Wisconsin

On March 13th, 2020, Wisconsin Governor Tony Evers ordered that all schools be closed indefinitely [3], as the threat of cases across the state was rising. In the next few months, counties adopted their own administrative orders in conjunction with CDC guidelines and statements from the Governor [4-6]. While not every county took these actions at the exact same time, most counties had similar orders in place at similar times, largely due to the counties’ willingness to comply with national level orders.

In May, the Wisconsin Department of Health Services (DHS) was reporting a 14-day downward trajectory in the percent positive rate of COVID-19 tests. This led to the continued re-opening of public places and small businesses, most of them operating at a limited capacity and still abiding the social distancing guidelines set previously [7]. While the DHS secretary warned about a spike in coronavirus cases due to this re-opening that hypothesis was not confirmed until mid-July, when a new seven-day average positive test records were set four times between July 15th and July 21st, 2020 [8-9]. Mask mandates were re-issued across the state [10] as a direct result of this increase in cases. Places that had made their guidelines more lenient for a brief period returned to a stricter set of rules.

Hospitalizations and deaths from COVID-19 continued to increase nationally through summer 2020. By late August, however, the Wisconsin Department of Health Services was reporting a new low for daily average new case count. As the total death toll in the state of Wisconsin reached 1000, the low daily average previously reported was followed by a string of record-highs across the board within just a few weeks. In late September Wisconsin was ranked the 4th-highest in the United States for total cases [11-12]. A graph of cumulative cases and deaths in Milwaukee County is shown in Figure 1 below:

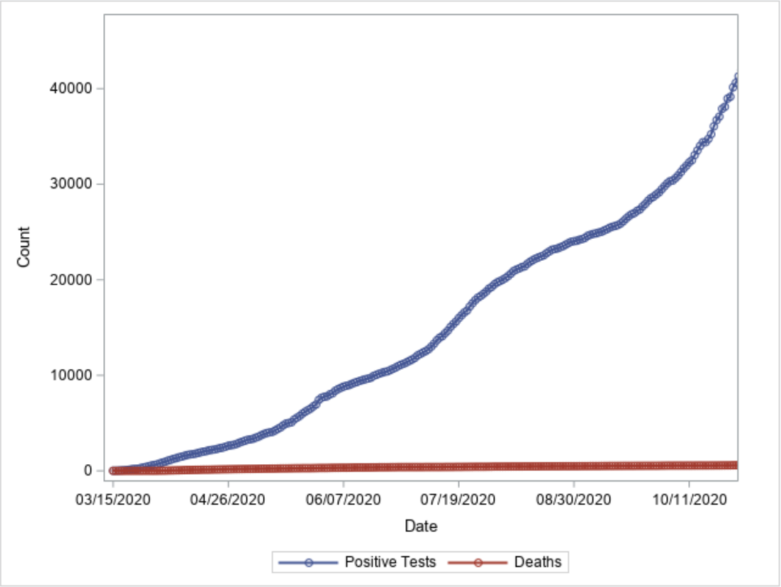

Figure 1. Cumulative confirmed positive COVID-19 cases and deaths in Milwaukee County

With these surges up and down in cases only at the *state* level, much less the national and international levels, there can be many confusing and conflicting inferences made. This prompts many questions and discussion from public health officials, epidemiologists, and biostatisticians alike. Some such questions are: What demographics are affected disproportionately by the coronavirus, with respect to infection and/or adverse health outcomes once infected? Given the trajectory of the epidemic curve, what is the predicted death count or mortality rate of the coming months? Have policy changes affected the likelihood of infection in any way?

Introduction to the Current Work

Due to the recent nature of these research questions, there are a few studies that have yet to be carefully reviewed and published, which adds an element of difficulty to literature searching. Nonetheless, there have been many efforts from groups of epidemiologists and biostatisticians to answer these questions to date. The critical points that separate these study approaches are the methods for collecting data and the specific statistical analysis carried out. Shinde et. al. outlines the current efforts to produce forecasting models for COVID-19, and divides them into four categories: Big Data, Social Media / Other Communication Data, Stochastic Theory and Mathematical Models, and Data Science / Machine Learning [13]. Within these four categories, these research questions, as well as many others in specific contexts, have been explored.

Logistic regression has been used in various contexts with COVID-19 data. Zhou et al [14] used three blood biomarkers in a logistic regression model to predict fatality of COVID-19 hospital patients. The study produced 96 true negatives and 12 true positives and predicted this with an average of 11.30 days in advance. Wang et al. used logistic regression to project deaths across the world with epidemiological data [15]. Li Yan et al. created an interpretable logistic regression model for COVID-19 patients, connecting several blood biomarkers to death [16] A cumulative logit model was used in a retrospective cohort study to determine the severity of COVID-19 cases [17]. A compendium of forecasting models exists to most exactly project infection and death rate up to two weeks in advance [18].

Statement of Hypotheses

With respect to Milwaukee County, the growing total of COVID-19 cases and deaths requires careful insight and inferences that could be made while modelling the risk of death, given a confirmed coronavirus case, adjusted for demographic variables. Due to the nature of the data, the risk of death given infection can only be examined at the univariate level, but that is still helpful. The following hypotheses will be tested when completing logistic regression on the dataset:

1. The risk of death for confirmed COVID-19 cases will be significantly higher for older age groups than for children or young adults.

2. The risk of death for confirmed COVID-19 cases will be significantly higher for black people than for any other race.
3. The risk of death for confirmed COVID-19 cases be equal for males, females, and cases of unknown gender.
4. The risk of death for confirmed COVID-19 cases, with respect to the demographic variables in question, will not be equal across the duration of time the dataset spans.

Hypotheses 1, 2, and 3 will be examined by building a cumulative logit model with respect to each variable individually, due to the nature of the data. Hypothesis 4 will be examined by comparing the odds ratios at each point in time across the entire duration of the dataset. It is from the results of answering Hypothesis 4 that inferences

DESCRIPTION OF THE DATA

Data Source

The dataset was acquired from the Wisconsin Department of Health Services on November 5th, 2020. It is a publicly available dataset that accepts reports of confirmed COVID-19 cases and deaths every day from hospitals and test centers throughout the state. The source of the data is the Wisconsin Electronic Disease Surveillance System (WEDSS). The case definition for COVID-19 is defined by the Centers for Disease Control and Prevention (CDC) and the Council of State and Territorial Epidemiologists. Although the reports of cases and deaths were accepted beginning on March 11th, 2020, demographic data was not included until May 11th, 2020. The dataset lists the age groups, races, and sex of confirmed COVID-19 cases and deaths on a cumulative basis as well as the confirmed COVID-19 cases and deaths overall on a daily and cumulative basis. These data are listed as column variables with the day of test/death as the row variable. For each demographic variable –

that is, age, sex, and race – the total number of confirmed COVID-19 cases and deaths is equal to the cases and deaths that day. Thus, from this data, there is no way to determine the *combination* of age, sex, and race, for a particular case or death on a particular day. Additionally, the negative test results were only reported with respect to the positive test results that day, and not with respect to the demographic variables.

Description of the Variables

The age variable was categorized by 10 years. Cases and deaths aged 30-39 years were the referent category for analysis, and the highest age group was 90 and older. The categories for the race variables are: “American Indian or Alaskan Native”, “Asian or Pacific Islander”, “African American or Black”, “White”, “Multiple Races or Other Race”, and “Unknown”. The categories for the sex variable are: “Male”, “Female”, and “Other”. Summary statistics for cases and deaths in each category are in Tables 1-3:

Race	**Positives (%)**	**Deaths (%)**	**Death Rate**
American-Indian / Alaska Native	331 (0.71)	0 (NA)	0.000
Asian	1702 (3.64)	20 (3.30)	1.175
White	24519 (52.4)	346 (57.1)	1.411
Black	10752 (23.0)	209 (34.5)	1.944
Multi/Other	5378 (11.5)	15 (2.48)	0.279
Unknown	4123 (8.81)	12 (1.98)	0.291
Total	**46805**	**606**	**1.295**

Table 1. Summaries of COVID-19 cases and deaths in Milwaukee County by race.

Sex	Positives (%)	Deaths (%)	Death Rate
Male	24956 (53.3)	295 (48.7)	1.182
Female	21738 (46.4)	310 (51.2)	1.426
Other	111 (0.24)	1 (0.17)	0.901
Total	**46805**	**606**	**1.295**

Table 2. Summaries of COVID-19 cases and deaths in Milwaukee County by sex.

Age Group	Positives (%)	Deaths (%)	Death Rate
9 and Under	2080 (4.44)	0 (NA)	0.000
10 – 19	4959 (10.6)	0 (NA)	0.000
20 – 29	11163 (23.9)	7 (1.16)	0.063
30 – 39	8719 (18.6)	6 (0.99)	0.069
40 – 49	6783 (14.5)	22 (3.63)	0.324
50 – 59	5767 (12.3)	45 (7.43)	0.780
60 – 69	3912 (8.36)	109 (18.0)	2.786
70 – 79	1920 (4.10)	155 (25.6)	8.073
80 – 89	1060 (2.26)	157 (25.9)	14.81
90 and Over	442 (0.94)	105 (17.3)	23.76
Total	**46805**	**606**	**1.295**

Table 3. Summaries of COVID-19 cases and deaths in Milwaukee County by age group.

METHODS

All regression equations, apart from analyzing the odds ratios over time to address Hypothesis #4, use the cumulative logit model. This model is a special application of logistic regression. The general equation for logistic regression is:

$$log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

Where π(x) is the likelihood of success. The random component of the proportion of success to failure in x follows a binomial distribution; therefore, traditionally, logistic regression is used for a binary outcome [19]. It is often used in survival analysis to model the likelihood of death. However, due to the constraints of the dataset, the proportion of those recovered from a confirmed COVID-19 case as well as the proportion of those who took tests and tested negative were unavailable. The data was gathered to compare the new positive tests that accrued for each demographic over the past two weeks with the new deaths that accrued for that day and two weeks after. Several reports have placed the time from infection to death anywhere from 8 days to 8 weeks [19-21] although several estimates could be based on the time of inference. Some have predicted the time from infection to death will increase over time, because a more resilient population is becoming infected and fewer are likely to have a high-risk confirmed COVID-19 case [22]. Nonetheless, from those gathered values, a survival rate was calculated.

Although the prediction for x in logistic regression can be for a rate, since it is between 0 and 1 and follows the restrictions of logistic models, it is not best practice [23]. Therefore, the survival rates calculated for the dataset were categorized based on their percentile rank and are presented as “Risk of COVID-19 Death”. If the survival rate was in the lower percentiles, it was categorized as “Very High” [Risk of Death], and “High”, “Medium”, and “Low” categories completed the response variable. Since different variables had no way of being compared, the percentiles varied slightly for each model, as the survival rate distributions varied within those as well.

Logistic Regression for Each Variable

The dataset had cumulative cases and totals with respect to each category. For every day with *n* total positives or deaths, the number of positives or deaths for each race, age group, and sex were *all* equal to *n*. Thus, separate cumulative logit models were used for each variable. The cumulative logit model shows the cumulative probability that a positive COVID-19 case of a particular age, race, or sex, is classified in Y category (“Very High”, “High”, “Medium”, “Low”) or *lower*. The general form of this cumulative probability is:

$$P(Y \leq j) = \pi_1 + \pi_2 + \cdots + \pi_j, j = 1, \ldots, J$$

Where J (uppercase) is the total number of categories (in this case, 4), and j (lowercase) is the specific category for which we calculate probability. For the entire model, the value of x_j is 1 if the COVID-19 case belongs to that group, and 0 if it is not. The logits of these probabilities are:

$$logit[P(Y \leq j)] = \log\left[\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right] = \log\left[\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}\right], j = 1, \ldots, J - 1$$

The cumulative probability of a confirmed COVID-19 case being placed in the category of “Very High” risk or lower is 1, as it is the highest category possible. Each cumulative logit has its own intercept, but the values of the coefficients stay constant under the proportional odds assumption. Thus, the logits that will be examined are the cumulative probability for a confirmed COVID-19 case being classified as “Low”, “Medium or lower”, and “High or lower”. The beta coefficients in each model represent the increased or decreased log odds in a specific classification. The exponentiation of the coefficient is the odds ratio between the group in question and the reference group. The reference group for the age model was 30-39. The reference group for race was white. The reference group for sex was male.

Selection of Time Intervals for Inference

An odds ratio for COVID-19 death was calculated on each day of cumulative tests with respect to the same reference group used in the regression models. Those odds ratios were plotted using linear regression against time. By testing the significance of the slope of these odds ratios, if the relative odds for any category, compared to the referent group, changed significantly over time, that would imply that odds of death from COVID-19 were not proportional for the duration of the dataset. If there are significant slopes, then the cumulative logit will be modelled again but with respect to smaller intervals of time, which may increase the ability to classify or differentiate between categories of risk of death. From there, inferences could be made about the policy changes, if significant differences in odds ratios align with these changes or the ensuing public response.

Model Evaluation

A ratio of 20% test, 80% training was used to form each cumulative logistic model. The classification tables for each model were analyzed to see where the model performed best. A confidence interval of 95% was used to determine statistical significance.

RESULTS

Table 4 shows the percentile ranks assigned to the survival rates as divided by race:

Calculated Survival Rate	Approximate Percentile	Risk of Death Category
Rate ≤ 0.980	27	Very High
0.980 ≤ Rate ≤ 0.990	42	High
0.990 ≤ Rate ≤ 0.995	61	Medium
Rate > 0.995	100	Low

Table 4. Percentile ranks for corresponding risk of death category for race model.

Table 5 shows the estimates for the coefficients in the race cumulative logit model. Table 6 shows the estimates for the intercepts for each logit.

Race	Estimate	95% Confidence Interval	p
American Indian / Alaska Native	-18.3	NA	< 0.001
Asian	-0.007	(-0.415, 0.431)	0.97
White	(reference)	(reference)	(reference)
Black	0.032	(-0.368, 0.431)	0.88
Multi/Other	-1.773	(-2.218, -1.328)	< 0.001
Unknown	-2.084	(-2.570, -1.600)	<0.001

Table 5. Estimates for regression coefficients in race cumulative logit model.

Intercept (Pr(Y ≤ j))	Estimate	95% Confidence Interval	p
Low	-0.990	(-1.290, -0.690)	< 0.001
Medium or Better	0.163	(-0.122, 0.449)	0.35
High or Better	1.274	(0.960, 1.588)	< 0.001

Table 6. Estimates for intercepts in race cumulative logit model.

The percentile ranks for categorizing the risk of death by survival rate with respect to sex are shown below in Table 7:

Calculated Survival Rate	Approximate Percentile	Risk of Death Category
Rate ≤ 0.980	21	Very High
0.980 ≤ Rate ≤ 0.990	43	High
0.990 ≤ Rate ≤ 0.998	74	Medium
Rate > 0.998	100	Low

Table 7. Percentile ranks for corresponding risk of death category for sex model.

Table 8 shows the estimates for the coefficients in the sex cumulative logit model. Table 9 shows the estimates for the intercepts for each logit.

Sex	Estimate	95% Confidence Interval	p
Female	(reference)	(reference)	(reference)
Male	-0.370	(-0.803, 0.628)	0.09
Other	-19.4	NA	< 0.001

Table 8. Estimates for regression coefficients in sex cumulative logit model.

Intercept (Pr(Y ≤ j))	Estimate	95% Confidence Interval	p
Low	-1.789	(-2.183, -1.395)	< 0.001
Medium or Better	0.260	(-0.066, 0.586)	0.12
High or Better	2.801	(2.228, 3.374)	< 0.001

Table 9. Estimates for intercepts in sex cumulative logit model.

The percentile ranks for categorizing the risk of death by survival rate with respect to age are shown below in Table 10:

Calculated Survival Rate	Approximate Percentile	Risk of Death Category
Rate ≤ 0.900	25	Very High
0.900 ≤ Rate ≤ 0.960	34	High
0.960 ≤ Rate ≤ 0.990	47	Medium
Rate > 0.990	100	Low

Table 10. Percentile ranks for corresponding risk of death category for age model.

Table 11 shows the estimates for the coefficients by age group in the age cumulative logit model. Table 12 shows the estimates for the intercepts for each logit.

Age	Estimate	95% Confidence Interval	p
9 and Under	-8.835	(-52.4, 34.8)	0.69
10-19	-8.835	(-52.4, 34.8)	0.69
20-29	-0.275	(-1.136, 0.585)	0.53
30-39	(reference)	(reference)	(reference)
40-49	-1.226	(-2.374, -0.078)	0.035
50-59	1.131	(0.045, 1.816)	0.001
60-69	3.267	(2.618, 3.916)	< 0.001
70-79	4.050	(3.380, 4.720)	< 0.001
80-89	6.373	(5.632, 7.114)	< 0.001
90 and Over	6.420	(5.670, 7.171)	< 0.001

Table 11. Estimates for regression coefficients in age cumulative logit model.

Intercept (Pr(Y ≤ j))	Estimate	95% Confidence Interval	p
Low	2.333	(1.760, 2.904)	< 0.001
Medium or Better	4.099	(3.483, 4.715)	< 0.001
High or Better	5.578	(4.920, 6.236)	< 0.001

Table 12. Estimates for intercepts in age cumulative logit model.

The results of the linear regression of odds ratios for race, sex, and age respectively over time are shown in Tables 13-15. These linear regressions were all done at the univariate level (i.e., odds ratios of Black to White were regressed on time, as was odds ratios of Asian to White, American Indian/Alaska Native to White, etc. for all variables).

Race	Slope	95% Confidence Interval	p
American Indian / Alaska Native	0	0	NA
Asian	4.09e-04	(-0.004, 0.005)	0.86
White	(reference)	(reference)	(reference)
Black	7.03e-04	(-0.001, 0.003)	0.48
Multi/Other	-2.67e-03	(-4.55e-03, -9.73e-04)	0.003
Unknown	-5.85e-04	(-0.005, 0.003)	0.77

Table 13. Estimates for regression of race death odds ratio on time.

Sex	Slope	95% Confidence Interval	p
Male	(reference)	(reference)	(reference)
Female	1.53e-03	(1.73e-04, 2.89e-03)	0.03
Other	0	0	NA

Table 14. Estimates for regression of sex death odds ratio on time.

Age	Slope	95% Confidence Interval	p
Under 9	0	0	NA
10-19	0	0	NA
20-29	-0.004	(-8.28e-03, -1.50e-04)	0.04
30-39	(reference)	(reference)	(reference)
40-49	4.39e-18	(-6.12e-19, 9.39e-18)	0.09
50-59	3.40e-03	(-4.96e-03, 0.012)	0.423
60-69	-0.031	(-0.051, -0.010)	0.004
70-79	-0.030	(-0.382, 0.321)	0.87
80-89	0.343	(0.228, 0.457)	< 0.001
90 and Over	0.718	(-0.398, 1.833)	0.21

Table 15. Estimates for regression of age death odds ratio on time.

DISCUSSION

The cumulative deaths from COVID-19 by race are overwhelmingly by white and black people, who combine for 91.6% of the deaths in Milwaukee County. These two races comprise 75.4% of the cumulative confirmed COVID-19 cases, so that disparity in the proportions means that the Multi/Other and Unknown races' cases died at a much lower rate. The odds a black person's confirmed COVID-19 case being classified as "high" or better compared to white COVID-19 cases is approximately 3% higher. However, the parameter was not statistically significant in the model. The residual deviance of the model was 1690 (p < 0.001), suggesting that this model was not a good fit for the data. It completely misclassified the data, predicting all groups to have "Low" risk, per the misclassification table. The inferences made from this model, without separating into time intervals, are not ideal. This is because of the nature of the reference category as well as the other categories.

The survival rates for whites at the beginning of the timeline are ‘very high’, and the risk category goes down to ‘high’ and ‘medium’ in the months to follow. Several other races follow similar patterns, and this is likely due in part to the small difference in survival rate that were the cutoffs for each category of risk of death.

The cumulative deaths from COVID-19 by sex are fairly even – 295 deaths for females and 310 deaths for males. It holds logically, then, that the odds of a male being classified into a “High” risk of death category or higher are slightly higher than that of a female (or of Other gender, which experienced one death as an entire category). The interpretation of the cumulative logit model is a little difficult in this case. Because the coefficient is negative for females, that means that the relative odds of a positive COVID-19 test being classified at any risk of death is approximately 69%, or 30% lower than for males. This coefficient was not statistically significant (p = 0.09), which means that in general, the odds of classification into any risk category are equal. However, it is important to note that the p-value is close to the threshold of 0.05, and further study may be required to discern this trend from the national one.

The cumulative deaths from COVID-19 in Milwaukee County by age shows a staggering rate of death for cases among people 70 and older. Approximately 8% of people aged 70-79, 12% of people aged 80-89, and 24% of people aged 90 and Over died from COVID-19. The differences in odds of this subpopulation having a risk category for death assigned to their confirmed case are much higher than the reference group of 30-39 years old. In fact, the range is so large between all categories that it hinders this model’s ability to classify deaths in the older groups. The calculations for likelihood of classification into each category for ages 80-89 are below:

$$P(Y \leq low) = \frac{exp(\alpha_{low} + \beta_{80-89})}{[1 + exp(\alpha_{low} + \beta_{80-89})]} = \frac{exp(2.333 + 6.373)}{[1 + exp(2.333 + 6.373)]} = 0.99983$$

$$P(Y \leq medium) = \frac{exp(\alpha_{medium} + \beta_{80-89})}{[1 + exp(\alpha_{medium} + \beta_{80-89})]} = \frac{exp(4.099 + 6.373)}{[1 + exp(4.099 + 6.373)]} = 0.99997$$

$$P(Y \leq high) = \frac{exp(\alpha_{high} + \beta_{80-89})}{[1 + exp(\alpha_{high} + \beta_{80-89})]} = \frac{exp(5.578 + 6.373)}{[1 + exp(5.578 + 6.373)]} = 0.99999$$

This makes the probability of a confirmed case of an 80-89-year-old person being classified as "medium" risk alone is 0.014%, and for "high risk alone" approximately 0.002%. A reason for this level of misclassification could be the reference group; since the death rate of confirmed COVID-19 cases among those aged 30-39 was only 0.069%, it is 91% likely that a positive case would be classified as "low". The effect plot for this model is shown below in Figure 2. It shows the proportion of deaths for each category of age that were placed in each category of risk for death:

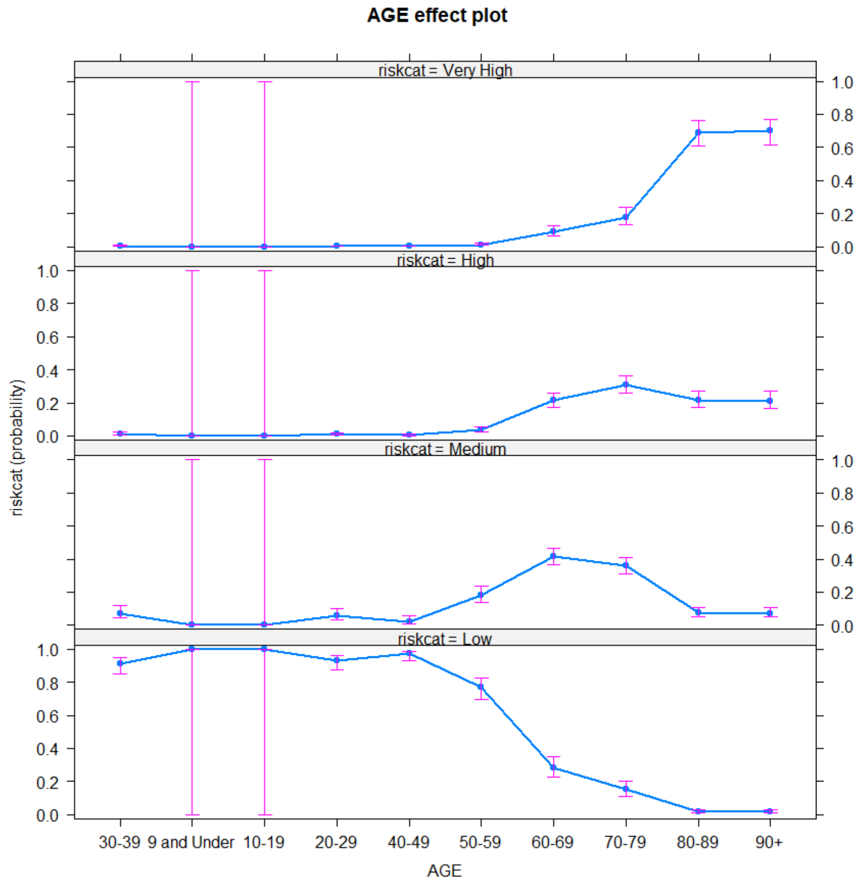

Figure 2. Effect plot of age groups' probability of death risk categorization.

The graph confirms the value of 91% for the reference category's placement in the "Low" risk category. It also shows that the older groups were almost completely unlikely to be in the "Medium" or "High" risk category. This highlights a discrepancy in the interpretation of the logit model. The calculation of probability for a confirmed COVID-19 case's death in the older categories made it seem like, almost certainly, every old categorization would be "low". However, Table 16 shows the classification table for the age model:

		Actual Values				
		Low	Medium	High	Very High	**Total**
Predicted Values	Low	202	14	0	0	**216**
	Medium	10	34	24	4	**72**
	High	0	0	0	0	**0**
	Very High	7	1	12	52	**72**
	Total	**219**	**49**	**36**	**56**	**360**

Table 16. Classification table of age cumulative logit model.

The model correctly classified 92% of "Low" categorized cases, 69% of "Medium" categorized cases, 0% of "High" categorized cases, and 93% of "Very High" categorized cases. Due to the phenomenon that was observed with calculating the probabilities of each coefficient, the difference between "Medium" and "High" categorization could have been too small for the model to properly differentiate. In fact, the model classified zero as "High" specifically, and seemed to split those values into "Medium" and "Very High".

It is important to note, however, that the model did well at classifying the extremes. Since the actual values for "High" were only 10% of the entire set of the responses, and they were mostly between consecutive classifications of "Medium" or "Very High", the model failed to separate these categories. This implies that the model would be more successful if proportional odds were not assumed, or if the dataset were broken into time intervals.

There may be no published or documented reason to think the assumption of proportional odds would not be met (with respect to demography, and certainly not comorbidities that would increase risk), but this discrepancy in the model indicates this may be the case. There are two solutions to this problem: 1) instead of a single β for a category of age across all risk categories, a different β_{low}, β_{medium}, β_{high}, and $\beta_{very\ high}$ for each age group, or 2) breaking up the dataset into time intervals that had a more constant risk, at least with respect to the reference group. The first solution would almost certainly require the proportions of negative tests or recovered patients per category to make any sort of valid inference. The second solution does not obfuscate the relative odds of death between age categories while providing a relative association between the initial stages and later stages of the virus' progression.

The linear regression of the categorical variables' odds ratios of death across time showed some significant results. In the race models, the Multi/Other category showed a significant slope of -0.00267. Although the value is not large in magnitude, it was statistically significant ($p = 0.003$). A negative slope would imply that the odds of death for those in the Multi/Other race, compared to the reference group of whites, decreased over time. A closer look at the dataset's classifications show that, indeed, the Multi/Other survival rate was maintained at 99% or above from September 17th to the end of the dataset.

Although the CDC's publication of race, ethnicity, and age trends in persons who died from COVID-19 [24] shows a disproportionate representation of Black and Hispanic people dying, this Milwaukee County data does not necessarily conclude that. The death rate for black people in Milwaukee County was about 1.9%, and the death rate for white people was about 1.4%. Milwaukee County is comprised of 545,872 white people and 251,870 black people [25]. This implies that approximately 4.5% of white people in Milwaukee County had a confirmed COVID-19 case, and 4.3% of black people had a confirmed COVID-19 case. This is contrary to multiple reports that typically focus on the cumulative case count that suggests black people are more susceptible to infection and death than white people, approximately 2-3 times the rate [26]. The issue with comparing this directly to current and previous studies, however, is that there is usually an adjustment for age, which is per our measurements as well the most powerful predictor of an adverse COVID-19-related health outcome.

The slope of the linear regression of female odds ratios on time was also significant (slope = 0.00153, p = 0.003). A positive slope implies that the odds of death for females increased over time, with respect to the reference group of males. This slope is more difficult to interpret. Consider the distribution of essential workers in different sectors with respect to sex [27]. Emergency services are comprised of 81% males, the transportation and delivery industry is 76% males, and industrial and commercial services are 86% males. These industries were affected by COVID-19 cases and deaths first, before administrative orders started protecting the essential workers. The healthcare sector, which is comprised of 76% females, was impacted by direct exposure to the virus later, during which time hospitals were reaching their maximum capacity. Thus, the relative odds of female high-risk cases increased.

The slope of the linear regression of age groups' odds ratios on time showed multiple significant variables. The 20-29 age group had a negative slope (-0.004, p = 0.04), as did the 60-69 age group (-0.031, p = 0.004). The 80-89 age group had a positive slope (0.343, p < 0.0001). Fortunately for the sake of interpretation, the reference age group survival rate was consistently "Low", except for an elevation to "Medium" from August 21st to September 3rd. The positive slope in the 80-89 age group is most likely due to lingering complications from COVID-19 that led to death outside of the 2-week window that was accounted for in this model. Since that could be a potential overestimation of deaths attributed to cases confirmed in October or November, a more exact method of tracing infection to death could help explain this relationship. The age group younger than the reference category with a negative slope could be explained by an increased resilience of that population over time, and people aged 20-29 with comorbidities that increase the risk for death contracting the virus first. The 60-69 age group's negative slope is the most challenging to interpret, and examination of the survival rate over time shows that the risk of death was categorized as "Low"37, "Medium" for 104, and "High" for 39 of the 180 time points. The concentration of "Low" categorizations was toward the end of the dataset, as did the reference group; thus, the concentrations of "High" categorizations must have either been at the beginning (i.e., an at-risk age category became slightly more exposed than their older counterparts because a proportion may still be working or interacting with people), or their relative odds for death was slightly lower toward the end of the dataset.

With this set of inferences in hand, breaking up the cumulative logit model into 2-month time intervals was practical for a proof-of-concept. As stated, this is not an ideal model, because we do not have access to negative test results, and we cannot merge the demographic variables and make inferences on more than one category at once.

The cumulative logit model applied to months May and June showed negative coefficients for ages 10-19, 40-49, and 60-69. The model placed every group into either "Medium" or "Very High", which correctly classified 77 of 100 datapoints. The residual deviance was 803, and the AIC was 827. The same strategy for months July and August resulted in the sign of the 60-69 coefficient switching to positive. This shows that in the first few months, the risk of death was low for this category, and as the cumulative prevalence of the disease increased, more people from this age group died with respect to the more resilient reference category. The model, again, placed every group into either “Medium” or “Very High”, correctly classifying 85 of 120 datapoints. The residual deviance was 836, and the AIC was 860. In October and November, the model managed to correctly classify 9 of the 16 datapoints in the “High” category, which is a vast improvement from the previous two iterations. The residual deviance and AIC were significantly reduced—to 581 and 605, respectively. It was here that the sign of the coefficient for ages 40-49 switched from negative to positive, though the value of the coefficient itself was not statistically significant (β = 0.45436, 95% CI = (-0.524, 1.433), p = 0.36). The increased performance of this model, as was shown by breaking up the dataset into time intervals, is likely due to the more relatively constant survival rate sustained by the reference group in these months. The risk category for death of 30-39-year-olds' confirmed COVID-19 cases was “Medium” for the first few days of September, and then “Low” (or even none) for the remainder. Since proportional odds are assumed in these models, that helps contribute to model robustness.

An option in building the model could have been to include time as a covariate; that is, add a multiplicative effect corresponding to when the COVID-19 case was confirmed. There are several issues that preclude us from doing that in this dataset. The first is that tests are not completed and reported in the real time of acquiring the virus. There may be several days between a case contracting COVID-19 and then confirming it at a testing site. These days could stretch longer for populations who cannot get themselves to a testing site, such as the extreme ends of age, or people who have had to work despite experiencing symptoms.

To that point, the *willingness* to go get tested and the *availability* of testing sites are two other caveats when considering time as a variable. For those unwilling to get tested, their case may have finally been confirmed upon, for example, arriving at a hospital – which would overestimate the number of cases later. Similarly, in inner cities with densely populated communities, the potential demand for testing could have not met supply, which would also underestimate the present case count. Contrary to the initial belief that population density was a driving force of COVID-19 infection and severity, a study by Hamidi et al [28] shows that it might be more closely related to the size of the metropolitan area.

Ultimately, there could have been a few different strategies used to process and interpret this data, although all of them come with separate caveats. Machine learning has shown success with classifying both binary and ordinal outcomes related to COVID-19 [29-31], but those are in smaller datasets and often have hospital data from discharged patients as well as those who died (i.e., not using aggregated data from testing sites). In principle, it makes sense than a random forest classifier or simple neural network could more efficiently characterize a singular person's risk for death given several demographic (and, hopefully, comorbid) characteristics than a cumulative logit model. However, if the data had included the proportion of positive to negative tests, or a proportion of people recovered, the model could have been more sophisticated and produced better classification rates.

LIMITATIONS

The two key limitations in making inferences based on this available data have to do with the nature of recorded positive COVID-19 tests and deaths:

1. The total number of positives in a particular day is spread across all age groups, followed by all sexes, followed by all races. That is, there is no way to measure, for instance, the risk of death given a confirmed COVID-19 case for a white male aged 30-39. The variables were aggregated separately, and thus, the variables had to be analyzed one at a time.

2. While the *total* number of negatives was reported via WEDSS/DHS, the number of negatives (and tests) with respect to *each demographic* was not recorded. Thus, there was no way to measure relative odds of a confirmed COVID-19 test adjusting for the number of tests a specific demographic has taken.

Both limitations are consequences of the data being aggregated from the surveillance system. If a smaller sample were picked – for example, case data from one of the testing sites or hospitals reporting to WEDSS – it is possible that individual case data could have been acquired. In most cases this most likely would not have been available due to privacy issues. In the interest of monitoring this problem with respect to easily discernible administrative orders, and making inferences thereafter, the county-level data from Milwaukee County was used.

The first limitation is less problematic than the second limitation with respect to the inferences that were made in this analysis. While it would be beneficial for a regression model or machine learning technique to predict the risk of a confirmed COVID-19 case given several demographic criteria, the limitation of aggregated data versus individual case data would have still been present. The inferences with respect to age, race, and sex individually still provide important information about the people who could be most at-risk.

The second limitation is more problematic because it essentially eliminates the use of traditional logistic regression with a binary response variable. If the negative test results were available for the same demographic variables of interest in the regression model, the log odds of positive test could be calculated. The odds of a positive test, as opposed to the odds of death or survival given a positive test, are arguably more important when it comes to making inferences about policy changes. The idea that this logistic regression is technically predicting a survival rate given that X people from J age group have become a confirmed case for COVID-19 does not detract from the inferences being made; rather, it is a caveat for using this exact model on future work.

There are a few additional limitations regarding the cases themselves and whether their outcomes align with what is considered death related to COVID-19. First, it is at least quantifiably possible that a confirmed COVID-19 case will survive longer than 14 days. Without the individual case data matching infection and, if applicable, death, an interval had to be picked to estimate deaths over a certain period. The cutoff point of 14 days was deemed a reasonable interval of time for acquiring a new death for at least one category of each of the variables analyzed. It also holds logically that, if the rate of survival were to increase or decrease within these intervals over time, inferences could be made about those changes irrespective of it being the truly observed interval from infection to death across all cases. Second, it is also plausible that a case could have died from COVID-19 without taking a test initially, making the time of death equal to the time of infection. It is assumed in completing this analysis that this is a much more special case than the first issue.

Regarding the assessment of the model's predictions, these models themselves cannot be 'proven' in the sense that we cannot confirm the specific locations of positive cases and deaths as they took place in certain areas over Milwaukee County. The inferences in the separate cumulative logit models (separated by 2 months' time) are limited, because we can only relate that to the demographic proportions of essential workers or those who may have been most affected by travel. One possible solution would be at the hospital level and the effective use of contact tracing. The census tract- (i.e., neighborhood-) level data for Milwaukee County demography is available; it could be useful to predict the relative risk or odds of infection of a certain neighborhood given the difference in odds between certain demographics and the proportion of those demographics within the census tract. The model would look like a Cox proportional hazards model, and a "risk" score could be calculated based on the prevalence of a comorbidity or how infected the census tract already is. If the model is effective in determining which specific areas have the greatest prevalence of infection or death, then that model could offer insight on the community's response to COVID-19 safety administrative orders.

CONCLUSION

The cumulative logit model provided helpful insights as to determining the odds of a high-risk confirmed COVID-19 case for residents of Milwaukee County across age, race, and sex. Age was the most significant predictor as far as classifying the cases correctly. Plotting the odds ratios across time showed that, in fact, the proportional odds assumption could not be met exactly – which explained the poor deviance/AIC scores from the models as well as high misclassification rates. While inferences could be made about how the odds of death changed within certain time intervals, it was not clear how those inferences could be traced back to policy changes; rather, the interpretation made more sense when comparing to the proportional demography of essential workers. The future work should account for the observation that the odds are not proportional, and whether this is due to the recovery rate in certain areas or confounding factors that could not be explored here, the rates of infection and death should be considered with respect to the virus' progression through the area of interest.

REFERENCES

1. https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/
2. https://www.dhs.wisconsin.gov/news/releases/020520.htm
3. https://www.usnews.com/news/best-states/wisconsin/articles/2020-03-13/elmbrook-schools-suspend-classroom-teaching-moving-online
4. Admin order PDF
5. Admin order PDF
6. Admin order PDF
7. https://content.govdelivery.com/accounts/WIGOV/bulletins/28b7302
8. https://wkow.com/2020/07/15/state-reports-821-new-covid-19-cases-fourth-highest-total-thus-far/
9. https://wkow.com/2020/07/20/318366/
10. https://urbanmilwaukee.com/2020/09/22/evers-declares-new-health-emergency-extends-mask-mandate/
11. https://www.usnews.com/news/best-states/wisconsin/articles/2020-09-27/wisconsin-tops-2-000-covid-19-cases-for-4th-straight-day
12. https://www.beckershospitalreview.com/public-health/states-ranked-by-confirmed-covid-19-cases-july-1.html
13. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7289234/pdf/42979_2020_Article_209.pdf
14. https://arxiv.org/abs/2006.16942
15. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7328553/
16. https://www.nature.com/articles/s42256-020-0180-7

17. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7369341/

18. https://www.aha.org/guidesreports/2020-04-09-compendium-models-predict-spread-covid-19

19. https://www.hsph.harvard.edu/news/hsph-in-the-news/data-animation-shows-time-lag-between-covid-19-cases-and-deaths/

20. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200301-sitrep-41-covid-19.pdf?sfvrsn=6768306d_2

21. https://www.thelancet.com/cms/10.1016/S1473-3099(20)30769-6/attachment/380c180c-279c-4388-8017-28f52606909c/mmc1.pdf

22. https://pubmed.ncbi.nlm.nih.gov/33104158/

23. Agresti

24. https://www.epi.org/blog/who-are-essential-workers-a-comprehensive-look-at-their-wages-demographics-and-unionization-rates/

25. http://www.healthcompassmilwaukee.org/index.php?module=DemographicData&controller=index&action=index

26. https://www.apmresearchlab.org/covid/deaths-by-race

27. https://www.tandfonline.com/doi/full/10.1080/01944363.2020.1777891

28. https://www.cdc.gov/mmwr/volumes/69/wr/mm6942e1.htm

29. Mehta et. Al, “Early Stage Machine-Learning Based Prediction of US County Vulnerability to the COVID-19 Pandemic: A Machine Learning Approach”
30. Fong et. al “Finding an accurate early forecasting model from small dataset” arXiv: 2003. 107762020
31. Kutmar J, Hembram KPSS. “Epidemiological study of novel Coronavirus 2020” arXiv: 2003.07347 2020.