Spring 5-5-2022

Factors Affecting Time to Recovery: A COVID-19 Survival Analysis

Fernanda Montoya
Z1881015@students.niu.edu

NORTHERN ILLINOIS UNIVERSITY

Factors Affecting Time to Recovery: A COVID-19 Survival Analysis

A Capstone Submitted to the

University Honors Program

In Partial Fulfillment of the

Requirements of the Baccalaureate Degree

With Honors

Department Of

Statistics and Actuarial Science

By

Fernanda Montoya

DeKalb, Illinois

May 2022

University Honors Program
Capstone Faculty Approval Page

Capstone Title (print or type)
"Factors Affecting Time to Recovery: A COVID-19 Survival Analysis"

Student Name (print or type)
Fernanda Montoya

Faculty Supervisor (print or type)
Dr. Haiming Zhou

Faculty Approval Signature 

Department of (print or type)
Statistics and Actuarial Science

Date of Approval (print or type) **5/2/2022**

Date and Venue of Presentation
Statistics Colloquium April 29th, 2022

Check if any of the following apply, and please tell us where and how it was published:

☐ Capstone has been published (Journal/Outlet):

_____

☐ Capstone has been submitted for publication (Journal/Outlet):

_____

Completed Honors Capstone projects may be used for student reference purposes, both electronically and in the Honors Capstone Library (CLB 110).

If you would like to opt out and not have this student's completed capstone used for reference purposes, please initial here: _____ (Faculty Supervisor)

HONORS CAPSTONE ABSTRACT

This project is focused on the recovery rates of patients diagnosed with COVID-19 after different clinical trial drug treatments. Data for the clinical trial studied was obtained from the National Institute of Allergy and Infectious Diseases for the primary purpose of a survival analysis on patient time to recovery under a placebo and therapeutic drug treatment. Specifically, patients in this clinical trial were randomly selected to receive remdesivir, an antiviral drug, in combination with a placebo or baricitinib, a janus kinase inhibitor drug. Cox PH models were used to identify how the different treatment drugs affect time to recovery and time to death, along with what other patient factors may have an association with both time to recovery and time to death. It was found that the combination of remdesivir and baricitinib yields a faster recovery time than using remdesivir alone for treatment. Research on clinical trial drugs is vital for discovering possible treatments for widespread diseases such as the COVID-19 pandemic.

Introduction

From the onset of coronavirus disease 2019 (COVID-19), it has become apparent how ill-equipped healthcare around the world was prepared to deal with a pandemic of this nature. Much difficulty has been had in preventing the spread of COVID-19 and accommodating the needs of patients diagnosed with the disease. Part of this difficulty can be attributed to a lack of treatment options and cures available for the disease. In order to determine the effects of different treatment options or cures, clinical trial testing must be performed before any treatment options can be released to the widespread public. Clinical trials deal with the testing of different treatment options that are thought to be viable options for helping those suffering from a disease. In clinical trials, what is of main interest is how a patient responds to the treatment, and how survival of the patient from the disease is impacted with treatment.

To investigate the effects of different factors on a patient's survival, a common statistical technique for patient data in the biomedical field is survival analysis. Survival analysis is a data analysis approach that can answer questions on the probability of an event occurring. Typically, a time to event variable is required for a survival analysis to be conducted, and it is common that the event chosen is patient death, although the particular event can be specified as something other than death. Survival analysis has the capability of determining if certain characteristics affect survival rates in a population. This technique is then of great use for data involving the same time to event outcome for different populations. In this study, we would like to investigate how different populations' survival rates of COVID-19 are impacted by patient clinical trial treatments and different patient characteristics.

In particular, of interest in this study is how a patient's recovery time is impacted not only by the usage of clinical trial drugs, but also by patient characteristics such as sex, age,

race/ethnicity, and comorbidities. It has been shown in other survival analyses that these characteristics can play a role in patient deaths related to COVID-19 (Lu et al., 2021; Salinas-Escudero et al., 2020; Thai et al., 2020). Our time to event variable will measure the time to recovery and our analysis will seek to identify what characteristics assist with patient recovery, particularly as it relates to different clinical trial drug treatments.

Data

The data in this research was obtained from the National Institute of Allergy and Infectious Diseases' (NIAID) sponsored clinical trial study titled "Adaptive COVID-19 Treatment Trial 2" (ACTT-2). The dataset contains patient level data for hospitalized adults infected with COVID-19. Patients in ACTT-2 were separated into a trial group and a placebo group to test the effects of therapeutic drugs on recovery time. This clinical trial specifically sought information on the advantages of using two therapeutic drugs, remdesivir and baricitinib, in conjunction versus using remdesivir alone. Remdesivir is an antiviral drug that has been previously authorized by the Food and Drug Administration as a treatment for COVID-19 after analyses done in NIAID's sponsored "Adaptive COVID-19 Treatment Trial 1" (FDA, 2020B; ACTT-1 Study Group, 2020). Baricitinib is a janus kinase inhibitor that helps to decrease the activity of an overactive immune system and is commonly used for patients with rheumatoid arthritis for its inflammation blocking properties (FDA, 2020A).

In total, 1,033 patients diagnosed with moderate or severe cases of COVID-19 were randomly split into a combination drug treatment group (515 patients) and a remdesivir plus placebo drug treatment group (518 patients). The data received contained information not only on patient treatment type, but also on other patient characteristics such as age, race, ethnicity,

sex, BMI, region, etc. and noted a patient's recovery, recovery time, along with death, and death time. In both events of death and recovery, the event was indicated as happening with a "0" and censorship, or the event not occurring, was indicated with a "1." This censorship variable is what will be primarily used when conducting model fitting to assess time to recovery and time to death relationships. A more detailed breakdown of patient data included in the ACTT-2 files used for this analysis is specified in Appendix Figure 1.

Methods

As the primary focus of this research is to investigate the relationships that exist between patient recovery and explanatory variables of interest, we must assess any relationships that exist using statistical modeling techniques. For our analysis, there are three main steps involved with fitting these statistical models. The first step is to fit Kaplan-Meier curves, which are created from non-parametric estimates of the survival function of a population. The general formula for Kaplan-Meier estimates, as described by Kleinbaum and Klein (2012), is shown as

$$\hat{S}(t_{(f)}) = \hat{S}(t_{(f-1)}) \times \widehat{Pr}(T > t_f | T \geq t_f)$$

where the authors describe the estimate calculated with "survival estimate for a previous failure time is multiplied by the conditional probability of surviving past the current failure time." Essentially, Kaplan-Meier curves generate survival probabilities and communicate to us visually how one group's survival compares to that of another group's. Kaplan-Meier curves can be tested for significance with the use of a log-rank test to determine if one group's survival is truly different from another group's, as will be discussed in the results of our analyses.

Following the creation of these Kaplan-Meier survival curves, statistical modeling of the data is completed with the use of the Cox Proportional Hazards model (Cox PH). The Cox PH model is a popular model used in survival analysis for its robust nature that can closely approximate correct results for a variety of different data. The model itself can assess the relationships of different covariates of interest with survival of an event by using a formula that involves a baseline hazard function of time and an exponential function of covariates of interest. The formula for the Cox PH model (Kleinbaum & Klein, 2012) is shown as:

$$h(t, \boldsymbol{X}) \;=\; h_0(t) exp\left\{\sum_{i=1}^{p} \beta_i X_i\right\}$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ are explanatory variables (covariates of interest) and $h_0(t)$ is the baseline hazard function of time. From the formula, we can see how the hazard at a given time for an individual is determined using explanatory variables of interest. To relate this formula to our survival analysis, the hazard function generated from this formula will communicate the likelihood that a patient will recover at a certain time, with explanatory variables such as treatment type, age, sex, etc. taken into account. An important thing to note about the Cox PH model is that it requires the satisfaction of the proportional hazards (PH) assumption. Checking that a model meets the required PH assumption is the next step that will be taken in our model fitting approach for this data.

For the PH assumption to be met, the hazard ratio must be independent of time or constant over time. This means that one individual's hazard rate is proportional to the hazard rate of another individual. This is mathematically written (Kleinbaum & Klein, 2012) as

$$\hat{h}(t\,, \boldsymbol{X}^*) \;=\; \hat{\theta}\hat{\boldsymbol{h}}(\boldsymbol{t}, \boldsymbol{X})$$

where $\mathbf{X}^*$ and $\mathbf{X}$ denote individuals and $\hat{\theta}$ is a proportionality constant independent of time. To determine that a fitted Cox PH model meets the PH assumption, options to verify the PH assumption include graphical visualization of a constant hazard rate, goodness of fit statistical testing methods, and creating an extended Cox PH model with time-dependent variables. To verify the PH assumption graphically, we can check that the hazard curves of different groups of an explanatory variable are proportional, or we can check that the log-log survival curves of the variable do not cross one another; if either of these plots fail to show this proportionality or parallelism, the PH assumption is not met. Additionally, goodness of fit statistical testing can be performed on the explanatory variables to test for PH assumption satisfaction. Many testing approaches exist that are based on Schoenfeld residuals. These tests check if the Schoenfeld residuals for a covariate are uncorrelated with time in the model. If the p-value for these tests is shown to be statistically significant, then the PH assumption is not met. Finally, if goodness of fit testing has proven a covariate to be correlated with time, an extended Cox PH model that accounts for this time-dependent variable can be fit, where the original variable is multiplied by a function of time.

For the statistical analysis conducted in this study, a mixture of both R, a statistical programming language, and SAS, a statistical software, were used to create Kaplan Meier curves, conduct model fitting, and check model assumptions. Particularly, the data obtained from NIAID was first cleaned in R to create new groups for the variables of "Race" and "Ethnicity." This data processing was performed to assist with downstream analysis and combined race groups aside from "White" to "Other" or "Unknown." Similarly, the ethnicity variable was combined into "Hispanic or Latino," "Not Hispanic or Latino," or "Unknown." Kaplan Meier curves for time to recovery and time to death for the two different treatment groups were also

created with R, using the "survival" package and survfit() function to create survival curves, along with the "survminer" package and the ggsurvplot() function to plot the created curves. The remaining analysis, including model fitting and model checking, was performed in SAS. Cox PH model fitting and assumption checking can be performed with the PROC PHREG procedure in SAS. This PROC PHREG procedure involves the use of the *model* statement for model fitting and the use of the *zph* command for goodness of fit testing on covariates.

Results

To follow the procedure for performing our survival analysis, the first step conducted was obtaining Kaplan-Meier curves for our time to recovery data. We are primarily interested in if the different treatment types, combination of drugs treatment or placebo treatment, has an effect on patient time to recovery. The Kaplan-Meier curve generated for time to recovery for the different treatment groups can be seen in Figure 1. From the Kaplan-Meier curves created, we can see that the placebo group (indicated in blue) has a higher survival curve than that of the combination group (indicated in red). While in a time to death model this would indicate that the placebo group is better at achieving survival, in the case of our data, this means that the placebo group is better at "surviving" recovery, indicating that the combination treatment group has faster recovery times. The log-rank test statistic is indicated from our curves to be 0.017 (lower left-hand corner), meaning the difference between the two Kaplan-Meier curves is significant and the patient treatment group does influence patient time to recovery.

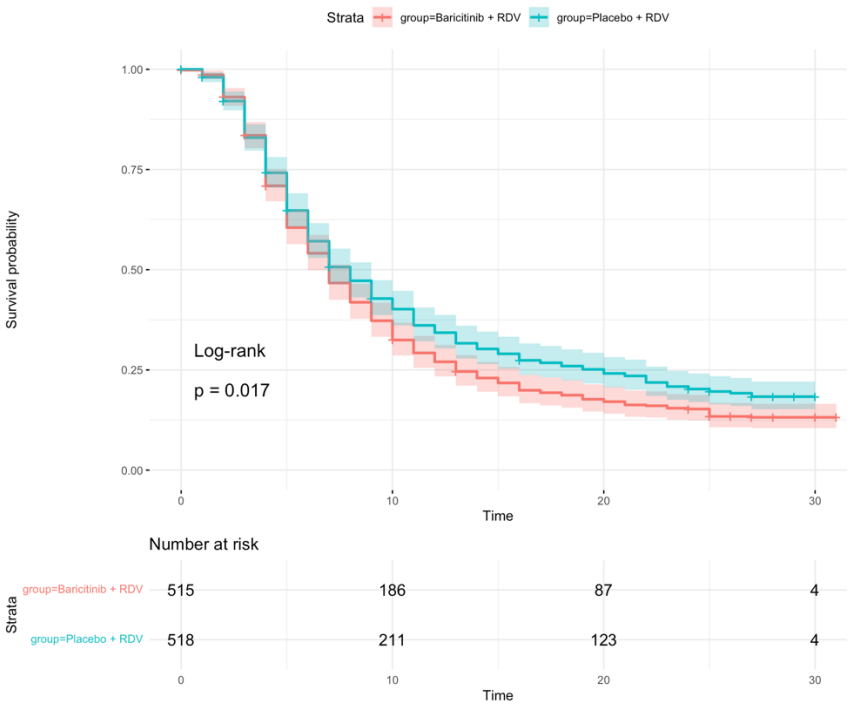

Figure 1. Kaplan-Meier curves for time to recovery for patients receiving combination drug treatment and those receiving placebo drug treatment.

After creating these Kaplan-Meier curves for our time to recovery data, we then want to proceed with fitting a Cox PH model to check the parameter estimate and hazard ratio for the treatment type effect. We fit a Cox PH model only including the treatment type covariate and test the goodness of fit with the *zph* command in SAS. The results of the *zph* command indicate that the treatment type variable is not significant (Appendix Figure 2), with a p-value of 0.3574. This means that we can proceed with the Cox PH model that was fit and analyze the maximum likelihood estimates generated from the model (Figure 2).

Parameter		Degrees Of Freedom	Parameter Estimate	Standard Error	Chi-Square	Pr > Chi-Square	Hazard Ratio
TRTP	Baricitinib + RDV	1	0.15777	0.06919	5.200	0.0226	1.171

Figure 2. Maximum Likelihood Estimates for Cox PH model including treatment type for time to recovery data.

From our table of the maximum likelihood estimates, we find the treatment type variable to be significant in our model (p-value 0.0226). The associated hazard ratio for this variable is reported as 1.171 and the parameter estimate for the variable is 0.15777. These values indicate to us that treatment type does have a significant effect on patient recovery time, particularly that patient recovery time is positively impacted by the combination of drugs. Those patients in the combination drug treatment group are associated with faster recovery times than those patients receiving remdesivir alone.

Continuing with our analysis after fitting this initial Cox PH model, we want to build off the model by adding in more covariates of interest. The covariates added to fit a new Cox PH model are treatment type, age, sex, race, BMI, ethnicity, region, baseline duration of symptoms, and a hypertension flag. Some covariates added to the model require a reference variable, which is what their hazard ratios are interpreted as being in comparison to. Particularly, the race, ethnicity, and region variables required reference groups which were, respectively, white patients, Hispanic or Latino patients, and the North America region. After adding these covariates to the model, we check the PH assumption with the zph command and find that the Asia region fails to meet the PH assumption for our model, indicated with a p-value less than 0.0001 (Appendix Figure 3). Failing to meet the PH assumption then requires us to go back to the Cox PH model and adjust it to ensure that the PH assumption is met.

To correct our model for the Asia region to meet the PH assumption, we fit an extended Cox PH model. This extended model will include a time-dependent variable for the covariate that fails the PH assumption, in our case this is the Asia region variable. An extended Cox PH model with a log-time dependent Asia region variable is then fit. This new variable will measure the interaction of the Asia region with time, thus satisfying the PH assumption. Proceeding with the Cox PH model maximum likelihood estimates, we then interpret the hazard ratio for both the Asia and Asiat variables as expressed as a function of time rather than the values that directly appear in the table (Figure 3). The hazard ratio for the Asia region can be expressed with this equation

$$HR = exp\{-3.93678 + 1.65690log(t)\}$$

that shows that as time increases in the study, the hazard ratio for the Asia region in comparison to the North America region increases with time. In looking at the remaining results in the table, the variables of treatment type, age, unknown ethnicity, and baseline duration of symptoms are all significant in the model. Age and unknown ethnicity were associated negatively with recovery time, while treatment type and baseline duration of symptoms are associated positively with recovery time.

Parameter		Degrees Of Freedom	Parameter Estimate	Standard Error	Chi-Square	Pr > Chi-Square	Hazard Ratio
TRTP	Baricitinib + RDV	1	0.15367	0.07125	4.6563	0.0309	1.166
AGE		1	-0.02049	0.00288	50.6064	<.0001	0.980
SEX	F	1	0.14424	0.07566	3.6345	0.0566	1.155
RACE3	OTHER	1	-0.07279	0.10521	0.4786	0.4890	0.930
RACE3	UNKNOWN	1	0.09429	0.09867	0.9132	0.3393	1.099
BMI		1	0.00175	0.00447	0.1530	0.6956	1.002
ETHNIC3	NOT HISPANIC OR LATINO	1	0.14596	0.09970	2.1433	0.1432	1.157
ETHNIC3	UNKNOWN	1	-0.68660	0.30987	4.9096	0.0267	0.503
Asia		1	-3.93678	0.67446	34.0693	<.0001	0.020
asiat		1	1.65690	0.28029	34.9433	<.0001	5.243
Europe		1	0.14261	0.29502	0.2337	0.6288	1.153
BDURSYMP		1	0.01863	0.00805	5.3566	0.0206	1.019
HYPFL	N	1	-0.00173	0.08021	0.0005	0.9828	0.998

Figure 3. Maximum Likelihood Estimates for time to recovery data with added covariates.

Having conducted the analysis for our initial interest of study with this data, we also are interested in analyzing the relationships that might exist between patient time to death and treatment type. The analysis for this time to death data will closely follow our analysis for the

time to recovery data. We begin with fitting Kaplan-Meier curves for the different treatment groups. In looking at the Kaplan-Meier curves created (Figure 4), we find that the combination drug group (indicated in red) does have a higher survival curve than the placebo group (indicated in blue). This would suggest a lower likelihood of dying with the combination treatment, but the log-rank test statistic indicates that the difference between the two survival curves is not significant. From the formal analysis, we cannot conclude that the time to a patient's death is affected by the treatment type received in the clinical trial.

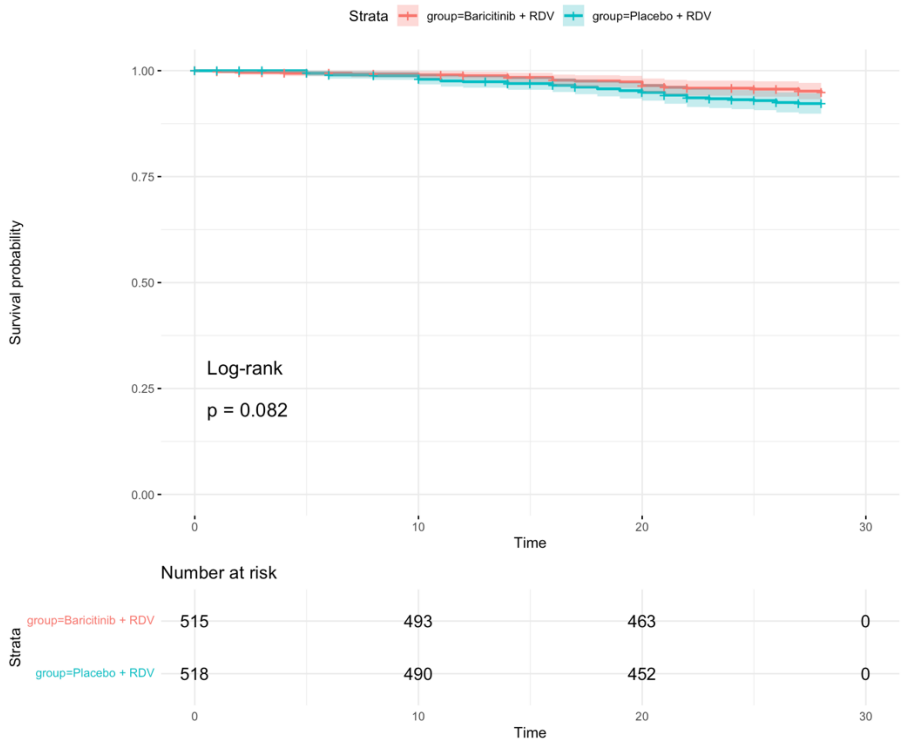

Figure 4. Kaplan-Meier curves for treatment type groups using time to death data.

After the creation of these Kaplan-Meier curves, we still proceed on with Cox PH model fitting. For our first time to death model, we include only the treatment type variable. The goodness of fit testing conducted on this model shows that the model meets the PH assumption (Appendix Figure 4). In looking at the maximum likelihood estimates for this model (Figure 5),

however, we note that just as in our Kaplan-Meier curves, treatment type is not significant and thus, treatment type is not associated with death for a patient.

Parameter		Degrees Of Freedom	Parameter Estimate	Standard Error	Chi-Square	Pr > Chi-Square	Hazard Ratio
TRTP	Baricitinib + RDV	1	-0.45177	0.26210	2.9709	0.0848	0.636

Figure 5. Maximum Likelihood Estimates for time to death data with treatment group variable.

Fitting a second Cox PH model with all covariates added that existed in the second model of our time to recovery data, we find that all covariates satisfy PH assumption testing (Appendix Figure 5). Then analyzing the table of maximum likelihood estimates for this model, we see that all of the covariates, aside from age, are insignificant in the model (Figure 6). Age being significant in the model and its positive parameter estimate of 0.05527 indicate that those patients with a higher age are associated with higher risk of dying. The remaining results of insignificant covariates tells us that other patient characteristics do not put them at a higher risk of death.

Parameter		Degrees Of Freedom	Parameter Estimate	Standard Error	Chi-Square	Pr > Chi-Square	Hazard Ratio
TRTP	Baricitinib + RDV	1	-0.34864	0.26668	1.7091	0.1911	0.706
AGE		1	0.05527	0.01208	20.9410	<.0001	1.057
SEX	F	1	0.02831	0.27914	0.0103	0.9192	1.029
RACE3	OTHER	1	-0.24083	0.36785	0.4286	0.5127	0.786

RACE3	UNKNOWN	1	-0.03174	0.37688	0.0071	0.9329	0.969
BMI		1	0.01321	0.01991	0.4399	0.5072	1.013
ETHNIC3	NOT HISPANIC OR LATINO	1	0.32633	0.35480	0.8460	0.7531	1.391
ETHNIC3	UNKNOWN	1	0.32994	1.04908	0.0989	0.7531	1.391
REGION	Asia	1	-1.24749	1.05043	1.4104	0.2350	0.287
REGION	Europe	1	-12.72828	674.27595	0.0004	0.9849	0.000
BDURSYMP		1	-0.00214	0.02854	0.0056	0.9401	0.998
HYPFL	N	1	-0.03115	0.30441	0.0105	0.9185	0.969

Figure 6. Maximum Likelihood Estimates for time to death data with added covariates.

Discussion

After this analysis, a few results from the data are of interest. To begin with the time to recovery analysis, the biggest finding is that the combination of remdesivir and baricitinib is a superior treatment to remdesivir alone for improving recovery time among patients. This finding is a vital breakthrough in identifying possible treatments for COVID-19. Currently, very few treatment options or cures are available for the disease, which is why the widespread impact of the disease has had a great effect for so long in our society. Finding options for treatments with the usage of clinical trials and statistical analysis offers promising directions for combatting COVID-19 in the future.

Secondary findings of the time to recovery analysis include how age and unknown ethnicity lead to longer recovery times. For patients of older ages, their hazard ratio indicates that their recovery periods are longer than those of a younger age. Similarly, patients with unknown ethnicity also are associated with longer recovery times as compared to patients of Hispanic or Latino descent. Interestingly, the baseline duration of symptoms, how many days a patient had symptoms prior to entering the study and receiving treatment, is associated with shorter recovery times as the number of days increases. Also, as noted in our extended Cox PH model, the Asia region must be represented with a time-dependent variable in our model. This informs us that the hazard ratio for the Asia region changes over time, and particularly increases over the study period as compared to the North America region. From looking at our Cox PH model overall, we see that many of the variables do not have hazard ratios much higher or lower than 1, which indicates that the clinical trial found consistent results across the different patient demographics.

Then looking at the time to death analysis, most of our results indicate that a patient's time to death is not associated with the different variables included in our models. The only significant covariate for the time to death models was age, which indicated that a higher age is associated with a higher risk of dying. The lack of significant results in these Cox PH models does indicate something positive, however, which is that treatment types are not associated with patient death. Additionally, the other covariates in the model also imply that a risk of death is not increased for any particular group of patients.

Conclusion

The results of our analysis indicate exciting results: the testing of these clinical trial drug treatments is significant and a combination of drugs yields better recovery times for patients as

compared to a singular drug. This conclusion was also found in an analysis of the data performed by the clinical trial research group associated with NIAID (Kalil et al., 2021). The results of this project also highlight interesting relationships that exist between different covariates of our events of interest. It is these relationships that make survival analysis so valuable as they can communicate to us any vulnerable populations that exist and a society and any underlying factors that healthcare may not be taking into account. Having knowledge of these relationships is what allows for future work to be done in public health reform.

COVID-19 and how it has affected the world has revealed many disparities amongst different groups in society. Research on how different demographics are affected by the disease indicate that those of lower-income backgrounds and minority status are at increased risk of being infected and impacted by COVID-19 (Aldridge et al. 2020; Khanijahani, 2020; Khanijahani et al. 2020; Paul et al., 2021). This research shows areas of weakness in healthcare that should be fixed. Being able to directly show if these relationships exist between patient socioeconomic status and a patient's death from COVID-19 offers valuable insight into how society functions and how improvements could be made to better the lives of all in a society.

Unfortunately, obtaining data with the information on different social determinants of heath can be difficult to come by. Often, a patient's income level or other identity characteristics are not indicated in data taken at the patient level. There has been a push for healthcare data to consider this information (Khalatbari-Soltani et al, 2020; Rogawski et al., 2016). Including such information offers a more holistic approach to data analysis and again, gives us the possibility of identifying interesting relationships that may exist. This area of public health research can make a large impact on societal health and how healthcare is structured in the future.

References

ACTT-1 Study Group. (2020). Remdesivir for the treatment of Covid-19: final report. *N Engl J Med*, *383*(19), 1813-1826.

Aldridge, R. W., Lewer, D., Katikireddi, S. V., Mathur, R., Pathak, N., Burns, R., ... & Hayward, A. (2020). Black, Asian and Minority Ethnic groups in England are at increased risk of death from COVID-19: indirect standardisation of NHS mortality data. *Wellcome open research*, *5*.

FDA. (2020A). Coronavirus (COVID-19) Update: FDA Authorizes Drug Combination for Treatment of COVID-19. U.S. Food and Drug Administration. Retrieved from https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-authorizes-drug-combination-treatment-covid-19

FDA. (2020B). FDA Approves First Treatment for COVID-19. U.S. Food and Drug Administration. Retrieved from https://www.fda.gov/news-events/press-announcements/fda-approves-first-treatment-covid-19

Kalil, A. C., Patterson, T. F., Mehta, A. K., Tomashek, K. M., Wolfe, C. R., Ghazaryan, V., ... & Beigel, J. H. (2021). Baricitinib plus remdesivir for hospitalized adults with Covid-19. New England Journal of Medicine, 384(9), 795-807.

Khalatbari-Soltani, S., Cumming, R. C., Delpierre, C., & Kelly-Irving, M. (2020). Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards. *J Epidemiol Community Health*, *74*(8), 620-623.

Khanijahani, A. (2021). Racial, ethnic, and socioeconomic disparities in confirmed COVID-19 cases and deaths in the United States: a county-level analysis as of November 2020. *Ethnicity & health*, *26*(1), 22-35.

Khanijahani, A., Iezadi, S., Gholipour, K., Azami-Aghdash, S., & Naghibi, D. (2021). A systematic review of racial/ethnic and socioeconomic disparities in COVID-19. *International journal for equity in health*, *20*(1), 1-30.

Kleinbaum, D. G., & Klein, M. (2012). *Survival analysis: A Self-Learning Text* (3rd ed.). New York: Springer.

Lu, W., Yu, S., Liu, H., Suo, L., Tang, K., Hu, J., ... & Hu, K. (2021). Survival analysis and risk factors in COVID-19 patients. *Disaster Medicine and Public Health Preparedness*, 1-6.

Paul, A., Englert, P., & Varga, M. (2021). Socio-economic disparities and COVID-19 in the USA. *Journal of Physics: Complexity*, *2*(3), 035017.

Rogawski, E. T., Gray, C. L., & Poole, C. (2016). An argument for renewed focus on epidemiology for public health. *Annals of epidemiology*, *26*(10), 729-733.

Salinas-Escudero, G., Carrillo-Vega, M. F., Granados-García, V., Martínez-Valverde, S., Toledano-Toledano, F., & Garduño-Espinosa, J. (2020). A survival analysis of COVID-19 in the Mexican population. *BMC public health*, *20*(1), 1-8.

Thai, P. Q., Son, D. T., Van, H. T. H., Minh, L. N., Hung, L. X., Van Toan, N., ... & Khoa, N. T. (2020). Factors associated with the duration of hospitalisation among COVID-19 patients in Vietnam: A survival analysis. *Epidemiology & Infection*, *148*.

Appendix

Variable		Frequency	Percent	Cumulative Frequency	Cumulative Percent
RECCNSR	0	839	81.22	839	81.22
	1	194	18.78	1033	100.00
DTHCNSR	0	61	5.91	61	5.91
	1	972	94.09	1033	100.00
TRTP	Baricitinib + RDV	515	49.85	515	49.85
	Placebo + RDV	518	50.15	1033	100.00
SEX	F	381	36.88	381	36.88
	M	652	63.12	1033	100.00
RACE3	OTHER	278	26.91	278	26.91
	UNKNOWN	259	25.07	537	51.98
	WHITE	496	48.02	1033	100.00
ETHNIC3	HISPANIC OR LATINO	531	51.40	531	51.40
	NOT HISPANIC OR LATINO	486	47.05	1017	98.45
	UNKNOWN	16	1.55	1033	100.00

REGION	Asia	67	6.49	67	6.49
	Europe	13	1.26	80	7.74
	North America	953	92.26	1033	100.00

Appendix Figure 1. Frequency table for covariates in time to recovery and time to death models.

Transform	Predictor Variable	Correlation	ChiSquare	Pr > ChiSquare	t Value	Pr > |t|
RANK	TRTPBaricitinib_RDV	0.0318	0.8472	0.3574	0.92	0.3580

Appendix Figure 2. Goodness of fit *zph* testing results for treatment variable in time to recovery model.

Transform	Predictor Variable	Pr > ChiSquare
RANK	TRTPBaricitinib_RDV	0.2504
RANK	AGE	0.7333
RANK	SEXF	0.8552
RANK	RACE3OTHER	0.4957
RANK	RACE3UNKNOWN	0.0927
RANK	BMI	0.6658
RANK	ETHNIC3NOT_HISPANIC_OR_LATINO	0.5321
RANK	ETHNIC3UNKNOWN	0.4799
RANK	REGIONAsia	<.0001
RANK	REGIONEurope	0.8914
RANK	BDURSYMP	0.9789
RANK	HYPFLN	0.7860

Appendix Figure 3. Goodness of fit *zph* testing results for covariates of interest in time to recovery model.

Transform	Predictor Variable	Correlation	ChiSquare	Pr > ChiSquare	t Value	Pr > |t|
RANK	TRTPBaricitinib_RDV	0.0304	0.0564	0.8123	0.23	0.8160

Appendix Figure 4. Goodness of fit *zph* testing results

Transform	Predictor Variable	Pr > ChiSquare
RANK	TRTPBaricitinib_RDV	0.9687
RANK	AGE	0.6470
RANK	SEXF	0.7442
RANK	RACE3OTHER	0.8850
RANK	RACE3UNKNOWN	0.1179
RANK	BMI	0.1813
RANK	ETHNIC3NOT_HISPANIC_OR_LATINO	0.1539
RANK	ETHNIC3UNKNOWN	0.5237
RANK	REGIONAsia	0.1340
RANK	REGIONEurope	0.9998
RANK	BDURSYMP	0.2960
RANK	HYPFLN	0.9353

Appendix Figure 5. Goodness of fit *zph* testing results for covariates of interest in time to death model.

R CODE

```
### Survival Analysis Coding Framework ##

#Load packages
library(survminer) #for clean plotting
library(survival) #for survival analysis functions

#Load dataset and read in data
setwd("/Users/fernandamontoya/Desktop/ACTT2_Datasets/ACTT_2_original")
covidData <- read.csv(file="ACTT2.csv",)
head(covidData)

#Creating column for recovery event
covidData$EVENTR <- ifelse(covidData$RECCNSR == 0, 1, 0)
#Column for death event
covidData$EVENTD <- ifelse(covidData$DTHCNSR == 0, 1, 0)
head(covidData)

#Create new column for three race categories (White, Other, Unknown)
covidData$RACE3 <- ifelse(covidData$RACE == "WHITE", "WHITE",
ifelse(covidData$RACE == "UNKNOWN", "UNKNOWN", "OTHER"))

#Create new column for three ethnicity categories (Hispanic or Latino, Not, Unknown)
covidData$ETHNIC3 <- ifelse(covidData$ETHNIC == "HISPANIC OR LATINO",
"HISPANIC OR LATINO", ifelse(covidData$ETHNIC == "NOT HISPANIC OR LATINO",
"NOT HISPANIC OR LATINO", "UNKNOWN"))

#export cleaned data to .csv file to use in SAS
write.csv(covidData,
"/Users/fernandamontoya/Desktop/ACTT2_Datasets/ACTT_2_original/ACTT2_clean.csv", na =
"", row.names=FALSE)

#Define variables
timeR <- covidData$TTRECOV
timeD <- covidData$TTDEATH
eventR <- covidData$EVENTR
eventD <- covidData$EVENTD
group <-covidData$TRTP

#Kaplan-Meier non-parametric analysis by treatment for RECOVERY
kmsurvival2 <- survfit(Surv(timeR,eventR) ~ group)
ggsurvplot(kmsurvival2, data=covidData, risk.table=TRUE, conf.int=TRUE, pval=TRUE,
pval.method=TRUE, ggtheme=theme_minimal())

#Kaplan-Meier non-parametric analysis by treatment for DEATH
```

```
kmsurvival3 <- survfit(Surv(timeD,eventD) ~ group)
ggsurvplot(kmsurvival3, data=covidData, risk.table=TRUE, conf.int=TRUE, pval=TRUE,
pval.method=TRUE, ggtheme=theme_minimal())
```

SAS CODE

```
libname mydata \\Client\C$\Users\ fernandamontoya\ Desktop\ACIT2 Datasets\ACIT 2
original\;

*read data;
data covid;
set mydata.actt2_clean;
run;

data covid2;
set covid;
Europe=0;
Asia=0;
if region='Asia' then Asia=1;
if region='Europe' then Europe=1;
run;

proc freq data=covid:
tables reccnsr dthcnsr sex race3 ETHNIC3 region trtp;
run;

*Step 1: model treatment w/o interaction*;
proc phreg data=covid zph;
class trtp;
model ttrecov*reccnsr(1)=trtp;
hazardratio "trt" trtp:
run;

*Step 2: final model of added covariates*;
proc phreg data=covid zph;
class trtp Sex Race3 (ref='WHITE') ETHNIC3 (ref='HISPANIC OR LATINO') region
(ref='North America') hypfl stratum;
model ttrecov*reccnsr (1) =trtp Age Sex race3 BMI ETHNIC3 Region bdursymp hypfl;
strata stratum;
hazardratio
"trt" trtp;
run;

* Step 3: Time to death model w/ treatment variable*
proc phreg data=covid zph:
class trtp;
```

```
model ttdeath*dthcnsr(1)=trtp:
hazardratio "trt" trtp;
run;

* Step 4: model for time to death with covariates *;
proc phreg data=covid zph:
class trtp sex race3 (ref='WHITE' ) ethnic3 (ref='HISPANIC OR LATINO') region (ref='North
America' ) hypfl stratum;
model ttdeath*dthcnsr (1)=trtp age sex race3 BMI ethnic3 region bdursymp hypfl;
strata stratum;
hazardratio "trt" trtp;
run:

/*Using log time for asia region (Extend Cox PH model)*/
proc phreg data=covid2 zph;
class trtp Sex Race3 (ref='WHITE') Ethnic3 (ref='HISPANIC OR LATINO') hypfl stratum;
model ttrecov*reccnsr (1) =trtp Age Sex race3 BMI Ethnic3 Asia asiat Europe bdursymp hypfl:
asiat=asia*log(ttrecov):
strata stratum:
hazardratio "trt" trtp;
run;
```