Caleb Fornshell
Covid-19 Project
5/15/2020

Research Question: What are the survival and hazard functions in patients that have been diagnosed with COVID-19, and what effect does the age, sex, and country of diagnosis have on the patient's survival? COVID-19 affects different groups of people in unique ways. Quantifying who is most at risk and how the disease progresses in different populations can provide valuable information in the treatment and preparation of the disease.

Data: Three datasets were used in this study. One dataset contained individual level details for confirmed COVID-19 cases in South Korea. Another dataset contained individual level details for confirmed cases in India, and the last dataset contains case information for confirmed cases in Florida. The South Korean and Indian datasets are available from Kaggle, and the Florida dataset is compiled from daily reports released by the Florida Department of Health. Only cases with non-missing information will be used. These datasets will be combined into a single dataset to allow for easier computation in further analyses.

South Korean Data: Link: https://www.kaggle.com/kimjihoo/coronavirusdataset#Case.csv
The South Korean data comes from Kaggle's Data Science for COVID-19(DS4C) dataset. The dataset was built using information gathered from the Korea Centers for Disease Control and Prevention(KCDC). This study will focus on cases that were confirmed on 02/20/2020 through 3/26/2020 for a study period of 35 days(5 weeks). This time period begins just prior to a surge of cases and contains the days with the most confirmed cases. The variables of interest from this dataset include, age, sex, confirmed_date, released_date, and deceased_date. In order to perform survival analysis on this data, basic data cleaning, such as data type conversion and observation selection, was performed as well as creating a survival time and censored status variables. The survival time was defined as the time, in days, from the confirmed date to either death, recovery, or the end of the study period. Cases that ended in recovery or that were still ongoing by the end of the study period were considered censored, and patients that died before the end of the study had an observed event. Censoring information is given by the "end.status" variable. Further details of the data cleaning and preparation can be seen in the R-code provided in Appendix B. A visualization of a portion of the cleaned data is shown in Figure 1 of Appendix A.

Indian Data: Link:www.kaggle.com/sudalairajkumar/covid19-in-india#IndividualDetails.csv
The Indian data was collected by volunteers who gather information from reading state bulletins and official handles and then compile their findings on www.covid19india.org. This study is using information from cases that were confirmed between 2/29/2020 and 4/4/2020(5 weeks). The variables of interest from the Indian dataset are age, gender, diagnosed_date, current_status, and status_change_date. Again, basic data cleaning including data type conversion, observation selection was performed. Using the last 3 variables, a survival time and a censoring variable were constructed where the survival time and censoring status are similar to the Korean dataset. Slight modifications to the age and gender variables were

made in order to match the format of the Korean dataset. Further details can be found the R-code provided in Appendix B. A visualization of a portion of the cleaned data is shown in Figure 2 of Appendix A.

Florida Data: Link: https://www.floridadisaster.org/covid19/covid-19-data-reports/
The data from Florida was compiled through daily reports released through the public health department of Florida. The Florida data consisted of cases that were confirmed between the dates of 3/13/2020 and 4/17/2020. This data was presented in two parts; One dataset for deaths and another dataset for cases. I then had to merge the two datasets using the dplyr package in R. This dataset contained the variables Age, Sex, Confirmed Date, and Death Date(if applicable). I then had to calculate the survival time from Confirmed Date to Death Date or the end of the study period. This data did not provide information regarding when or if a patient had recovered, so the censored observations were all considered to have still been under study at the end of the study period. This should not impact the analysis as complete dates were known for non-censored observations. A visualization of the Florida data is shown in Figure 3 of Appendix A

Exploratory Data Analysis: Relevant descriptive statistics for the variables of interest (sex, age.group, date of diagnosis, and survival time) were performed. A large difference in sample sizes for each location was seen with Florida having the largest sample size followed by South Korea and then India(24,554, 2360, and 819). Early analysis suggests differences in country, age, and sex. Full results, tables, and graphs are shown in Appendix A.

Gender:
South Korea, Florida, and India had varying proportions of females: 57%, 50%, and 25%, respectively. This shows a large difference in the makeup of each country's cases. A difference in mortatility is not obvious by observing the cross tabulation tables of end state by gender.(For complete results see Appendix A: Figure 4, Tables 1-3)

Age Group:
For all three locations, most patients were working aged adults aged 20-69. The Indian data has a smaller proportion of patients older than 70. This could suggest that South Korea and Florida are focusing tests on at-risk older populations, or it may reflect demographic differences between locations. Based on the cross tabulation tables, it appears that the elderly population dies at an increased rate.(For complete results see Appendix A: Figure 5, Tables 4-6)

Date of Diagnosis:
Analysis of the dates of diagnosis show that the outbreaks began at different times for each location. It appears that the outbreak was controlled by the end of South Korea's study period. The progress of the outbreaks in the other locations is not apparent. It is important to note the changing scale of the y-axis, as Florida has many more daily cases.(For complete results see Appendix A: Figure 6)

End Status and Survival Time:
Analysis of the end status of cases shows that for all locations a small proportion of cases

had ended in death by the end of the study with South Korea's death rate being the lowest(SK:1.8%, India:2.8%, FL:2.9%), although many cases were still ongoing. This shows that many of the cases were censored observations (hospitalized/isolated, released, or alive). Analyzing simple descriptive statistics of survival times, it appears that there may be a difference in the survival times between locations(Appendix A: Table 7). South Korea has the longest mean survival time with India having the shortest. This could be due to differences in healthcare or the individuals being tested. Only the sickest may be getting tested in India, whereas South Korea may be testing more less severe cases. (For complete results see Appendix A: Figures 7-8, Tables 1-7)

Univariate Survival Analysis and Inference I will begin by analyzing one variable at a time. I will be using the stratified k-sample log rank test to determine if there are differences between different levels of a particular variable while controlling for the other variables. I will be testing the hypotheses:

$$\begin{aligned} &H_0 : S_1(t) = .. = S_k(t) where\ k\ is\ the\ number\ of\ levels\ for\ a\ variable \\ &H_1 : S_1(t) \neq ... \neq S_k(t) for\ at\ one\ i \neq j\ in\ 1, ..., k \end{aligned} \tag{1}$$

for each of the variables country, sex, and age. The k-sample log-rank test is a non-parametric test with few assumptions to check. The only assumption that needs to be met is that censored observations are not at increased risk of death. This assumption is met as there is no evidence to suggest otherwise.

Along with performing the stratified log-rank tests, both the survival and hazard functions will be drawn to illustrate the differences in the levels of the variables. Piecewise Kaplan-Meier survival curves and 95% confident intervals will be drawn using the `survdiff()` function from the survival package. Smoothed hazard curves will be drawn using the `muhaz()` function from the muhaz package. Smooth survival curves based on these hazard calculations will be shown along with the KM curves.

First, I constructed overall survival and hazard curves. This is shown in Figure 1. This figure shows that overall a patient is most at risk of dying shortly after being diagnosed with COVID-19.

The first variable analyzed was country. The survival and hazard curves are shown in Figure 2. This figure clearly shows that there is a difference between survival curves depending on the country of diagnosis. It appears that South Korea has a survival curve that is stochastically longer than both Florida and India. The stratified log-rank statistic supports this claim. A Chi-Square test statistic of 42 (p-value=8E-10) indicates highly significant results(Appendix A: Table 8). Further analysis of the test shows that this is largely due to India having substantially more deaths and South Korea having fewer deaths than expected under the null hypothesis. Pairwise comparisons between countries show that South Korea had stochastically longer survival times than India and Florida, and India and Florida were not significantly different from each other(Appendix A: Table 9).

It is worth noting that the smooth survival curve constructed by the `muhaz()` function for India does not closely follow the KM survival curve. Instead, the smoothed curve indicates a survival function that would be much worse for patients. This could be due to the high number of Indian patients who died shortly after diagnosis as shown in the steep drop in the Indian KM survival curve shortly after t=0 but levels off by t=5.

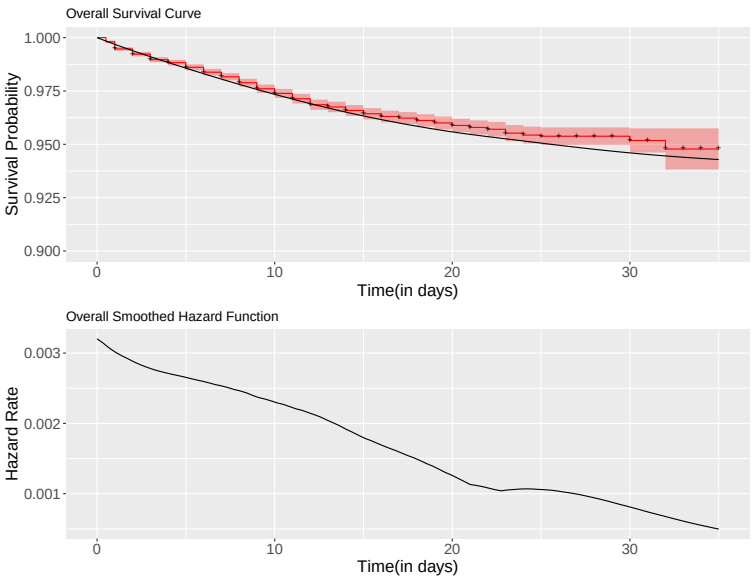

Figure 1: Overall Survival and Hazard Curve

One possible issue this analysis shows is that a proportional hazards model may not be appropriate. This is demonstrated as the survival curves of Florida and India intersect. In a proportional hazards model the survival curves should be parallel. This condition is examined further in a later section.

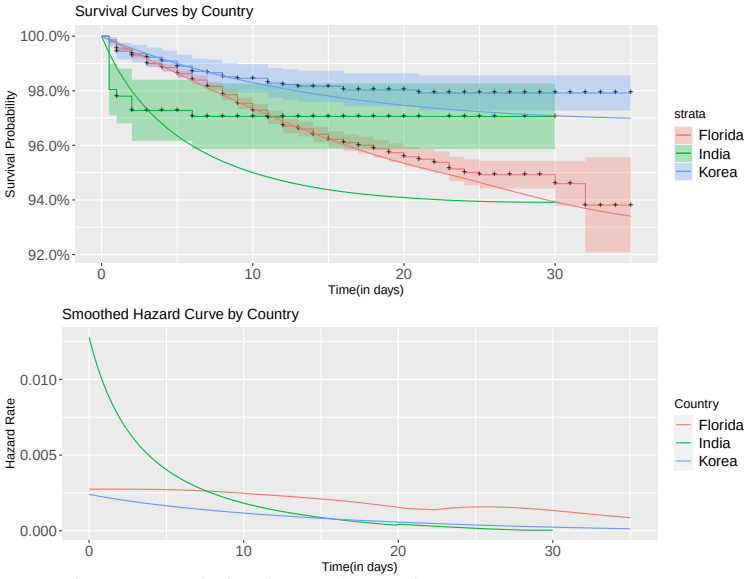

Figure 2: Survival and Hazard Curve by Country

The second variable examined was sex. The survival and hazard curves are shown in Figure 3. This figure shows that the survival curve for females is stochastically longer than for males. The stratified log-rank test also supports this claim(Appendix A: Table 10). The test statistic is a Chi-Square value of 44.3 (p-value=3E-11) indicating highly significant results. The table of results shows that females had 92 fewer deaths than would be expected if H_0 were true.

The last variable analyzed was age. I created a categorical variable which broke down the age into several categories, similar to those used in the exploratory data analysis. However, some groups with a low number of patients were combined to increase the sample size for groups and make the analysis more appropriate. The age groups used in completion of the stratified log-rank test were "<20", "30s", "40s", 50s","60s", and">70".

The survival and hazard curve for each age group are shown in Figure 4. This figure clearly shows that the older age groups are more at risk of dying from Covid-19 and have stochastically shorter survival curves. The four youngest age groups are nearly impossible to distinguish in this graph, but the older age groups each have a significantly lower survival

curve. The Chi-Square test statistic from the stratified log-rank test also supports the alternative hypothesis with a value of 1756 (p-value < 2E-16) indicating extremely significant results(Appendix A: Table 11).

Upon further analysis of the results we see that there were no observed deaths in the youngest age group and only one death in the "20s" group. We also see that the only age groups having a higher number of deaths than expected is the "60s" and ">70" groups with the ">70" group accounting for 70% of all deaths even though it only accounted for 17% of cases. Pairwise comparison between different age groups show that the two youngest groups(<20 and 20s) are not significantly different from each other, but are stochastically longer than all the other groups. We also see that the oldest group(>70) is stochastically shorter than all other groups. Complete pairwise comparisons are shown in Appendix A(Table 12), but in general, younger groups have longer survival times than older groups.

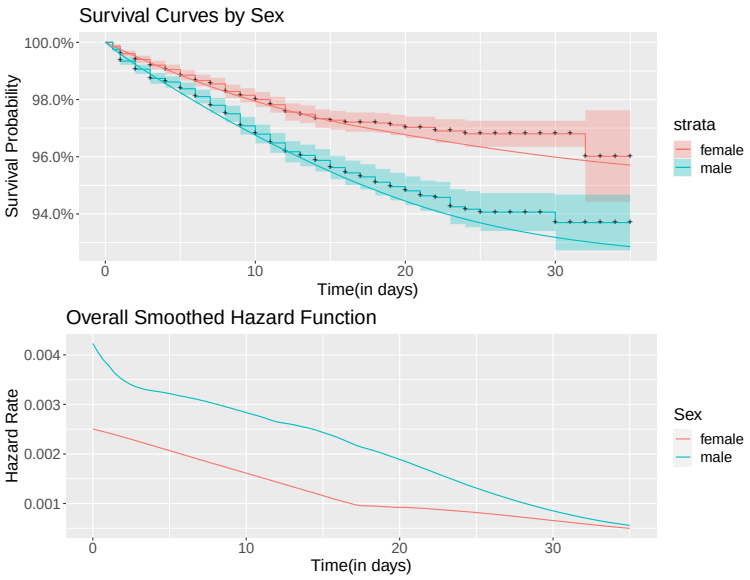

Figure 3: Survival and Hazard Curves by Sex

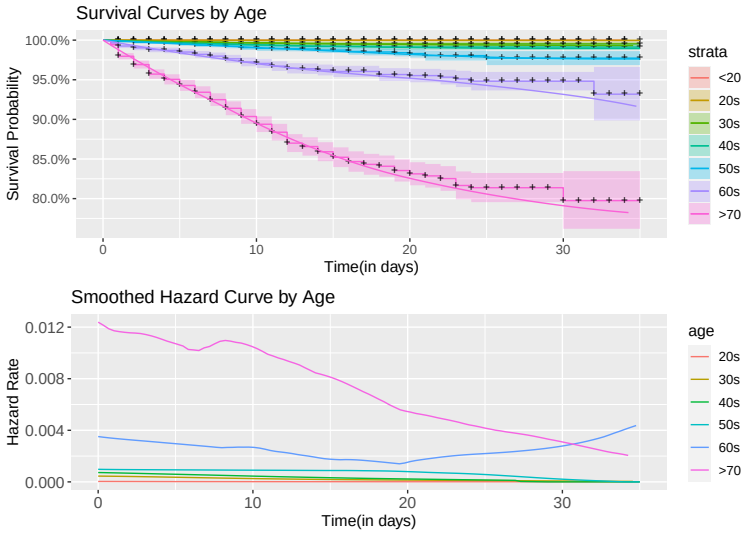

Figure 4: Survival and Hazard Curves by Age Category

These univariate analyses demonstrate that each of the three variables analyzed is significant while controlling for the other two variables. The disease appears to affect elderly populations and males more so than it does the young and females. The country of diagnosis also appears to have an effect. There are most likely many other variables such as underlying diseases and overall health that can affect the outcome of patients; however, this analysis was limited by the data available.

Fitting Cox Proportional Hazards Models In the following section Cox Proportional Hazards models will be fit to the COVID-19 survival data. Numerous Cox PH models will be fit with different combinations of covariates. The first models will use the age groups from the stratified k-sample log-rank tests from the previous section and other models will

consider age as a numeric value; that is, the patient's age in years.

	chisq	df	p
age.group	4.46	6	0.615
sex	2.92	1	0.088
country	35.96	2	<.0001
GLOBAL	41.88	9	<.0001

Table 1: Results from PH test

Before any model can be fit, the assumption of proportional hazards must be valid. The proportional hazard assumption assumes that the hazard rate for one patient is proportional to the hazard rate of another patient and this proportion is constant over all times. The proportional hazards assumption can be checked visually and using a hypothesis test. Visually, this assumption can be checked by inspecting the hazard curves of different groups of the sampled population. If the proportional hazards assumption is valid the curves should be parallel and never cross, although this does not guarantee the assumption is valid.

	chisq	df	p
age.group	3.36	6	0.760
sex	2.92	1	0.110
GLOBAL	5.92	7	0.550

Table 2: Results from PH test stratified by country

As shown in Figure 2, the survival and hazard curves for country are clearly not parallel. Thus, the assummption of proportional hazards is violated. This conclusion is supported by performing a hypothesis test using the `cox.zph()` function from the survival package. This procedure produces a p-value of <.0001 meaning the null hypothesis of proportional hazards is violated due to the variable country(Table 1). To correct for this issue, the models will be stratified on the variable country. Testing for violations in the stratified model generates insignificant p-values for both sex and age.group; the assumption of proportional hazards is reasonably met in the stratified model(Table 2).

	coef	exp(coef)	se(coef)	z	p
age.group20s	11.53	1.01e5	672.1	0.017	0.986
age.group30s	14.08	1.30e6	672.1	0.021	0.983
age.group40s	14.67	2.35e6	672.1	0.022	0.983
age.group50s	15.29	4.38e6	672.1	0.023	0.982
age.group60s	16.34	1.25e7	672.1	0.024	0.981
age.group>70	17.72	4.97e7	672.1	0.026	0.979
sexmale	0.4930	1.637	0.0748	6.954	4.27e-11

Table 3: PH models with discrete age groups. LRT=1411, 7 df, p<.0001, 778 events, n=27733

Using Cox regression, a stratified model is fit to the data, and the results are shown in Table 3. The model has a LRT test statistic of 1411 with an overall p-value <.0001. The most significant variable based on the Wald test scores is sex which indicates being male increases a patient's hazard rate by 64% when the other variables are held constant. The age group coefficients are all insignificant based on the Wald tests, which is unexpected as the stratified log-rank test indicated age group was significant. Therefore, a partial log-likelihood

test was used to determine if age.group should remain in the model. The results, shown in Appendix A Table 13, indicate that age.group is significant and should remain in the model.

	coef	exp(coef)	se(coef)	z	p
age	0.0871	1.0910	0.0025	34.151	<2e-16
sex	0.5653	1.7601	0.0751	7.529	5.11e-14

Table 4: PH models with numeric age. LRT=1589, 2 df, p<.0001, 775 events, n=27411

A more interpretable model may be achieved if the variable for age is considered a continuous numeric variable rather than using discrete categories. A model is fit using the age of the patient in years, sex, and stratifing by country. Using a hypothesis test, we see that proportional hazards can reasonably be assumed(Appendix A: Table 14), and that the two variables produce a statistically significant model with all coefficients having significance based on the Wald test statistics(Table 4). This model is easily interpretated as each one year increase in age increases the hazard rate by 9% with sex held constant. We also see this model increases the hazard rate for males by 76% when age is held constant. This model had a slightly smaller sample size as the age of patients in South Korea was calculated based on their birth year, and some patients did not have birth year recorded.

We can compare the two models to determine which performs better. This comparison can be made using Akaike Information Criterion(AIC). The model with the smaller AIC is preferred. This is an imperfect comparison as the model with continuous age has slighlty fewer observations. The model with discrete age groups had an AIC of 13,344 versus an AIC of 13,097 for the model with continuous age. We conclude that the second model is slighly better.

	coef	exp(coef)	se(coef)	z	p
age	0.087	1.091	0.003	34.165	<.0001
sexmale	0.565	1.760	0.075	7.526	<.0001
I(India)	3.364	28.892	0.385	8.733	<.0001
I(Korea)	0.209	1.233	0.257	0.814	0.4155
I(India):s.time	-1.020	0.361	0.289	-3.534	0.0004
I(Korea):s.time	-0.099	0.906	0.033	-2.978	0.0029

Table 5: Reduced extended Cox model. LRT=1664, 6 df, p<.0001, 775 events, n=354399. I(.) is the indicator function.

Extended Cox Model The models from the previous section were all stratified on country. This made calculating a coefficient and performing inference on the country variable impossible. An extended Cox model with numeric age and using an interaction between country and survival time can make inference on the country variable possible. An extended Cox model will consider the coefficient value for each country a time dependent variable. This is reasonable as the plot of Shoenfeld residuals shows that the coefficient for country is not constant over time(Appendix A: Figure 9). This figure shows that the value of the country coeffient is higher at smaller values of t. Fitting a model with interactions between all covariates and survival time also provides another method of testing the proportional hazard assumption. Fitting this full model shows that age and sex do not violate the PH assumption(Appendix A: Table 15) agreeing with the previous findings.

Based on these results a model was constructed using numeric age, sex, and interaction terms between the country indicators and survival time variables with Florida being the baseline location.

The results of this model(Table 5) indicate all coefficients are significant at an alpha level of 0.05 with the exception of the Korea indicator variable. The single largest contributing varible appears to be whether the patient is diagnosed in India with an estimated coefficient value of 3.364. A visualization showing the hazard ratios for each varible along with 95% confidence intervals is given in Figure 5. The value for the India indicator is too large to be easily shown in the chart.

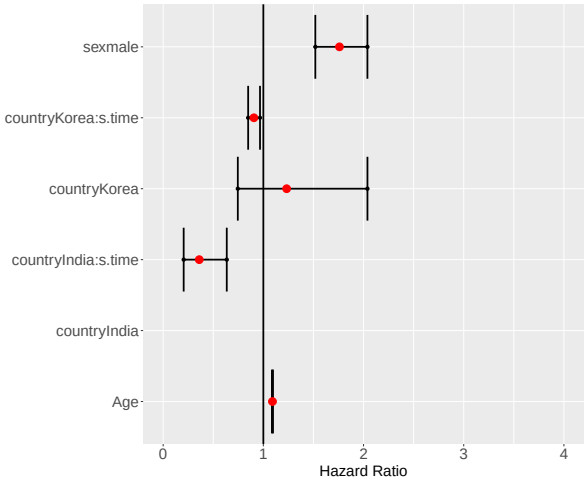

Figure 5: Hazard Ratios for variables in the simplified Extended Cox Model. I(India):s.time CI (13.58, 61.47)

Conclusions and Future Research This study demonstrates that age, sex, and country all have significant effects on the survival of COVID-19 patients. Performing stratified log-rank tests show that there are significant differences in the levels of all variables. In general, young patients and females are less likely to die from the disease. It also appears that of the three countries studied South Korean patients had the greatest chance of survival.

While this study provided significant results, there were a number of areas where the study could have been improved. Improvements could be made by including more factors that are thought to have an impact on a patient's survival. These factors could include the presence of underlying health conditions, the patient's race and economic status, the patient's treatments, and the inclusion of more locations. Many of these factors may be included as more time passes and more complete data is collected and made available to the public.

Another aspect of this analysis that could be improved and would add clarification to the results, would the inclusion of a variable indicating when the patient's first symptoms occurred. This information was available in a small number of South Korean cases, but not enough to include it in the analysis. Having this variable would allow for the progression of the disease to be more accurately described, and would allow for a more realistic hazard curves.

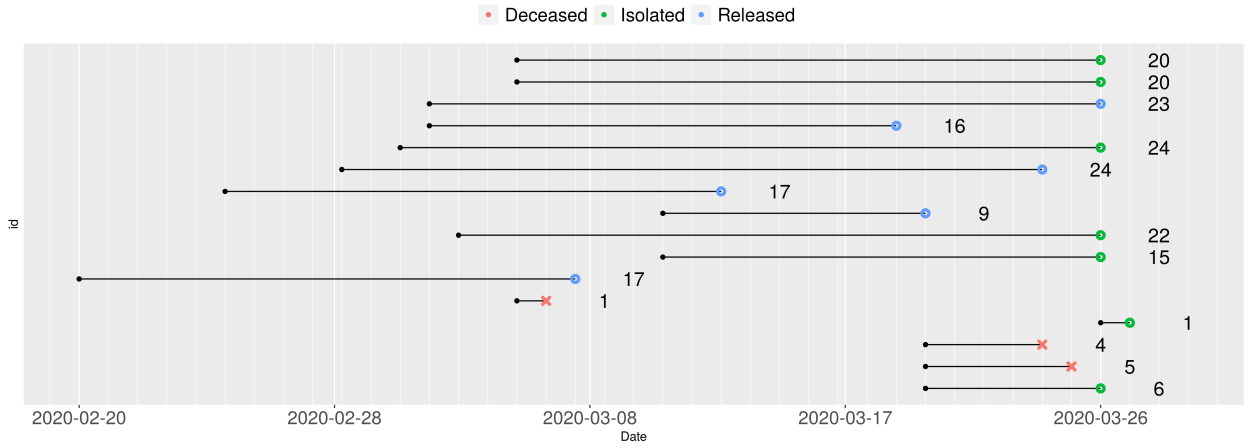

Figure 1: Visualization of South Korean Data. "X" represents a death was observed, and "O" represents a censored observation.

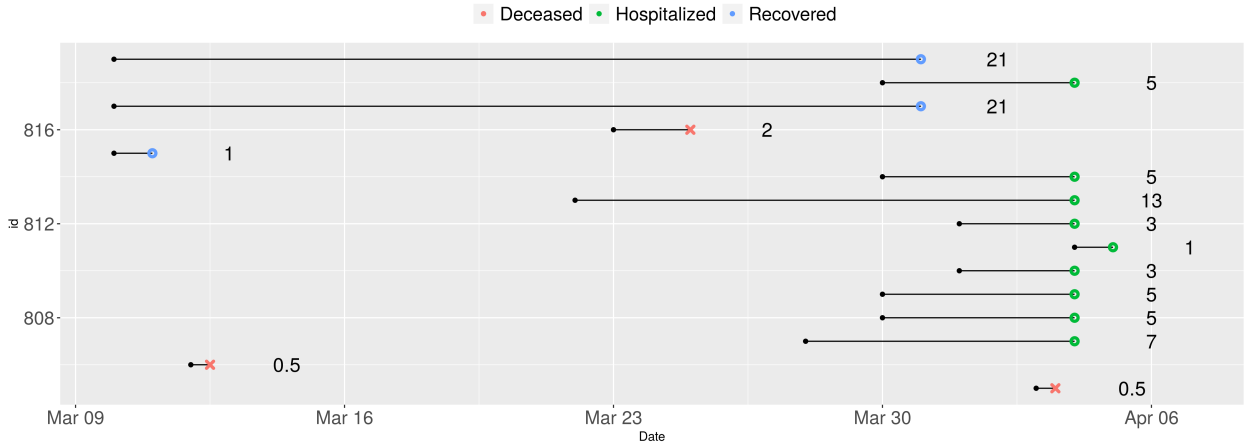

Figure 2: Visualization of Indian Data. "X" represents a death was observed, and "O" represents a censored observation.

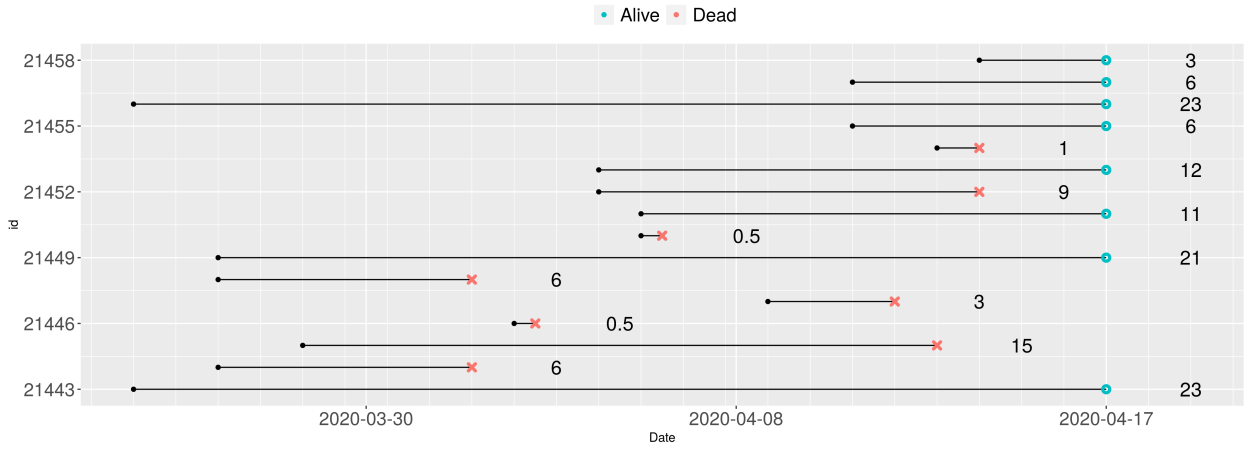

Figure 3: Visualization of Florida Data. "X" represents a death was observed, and "O" represents a censored observation.

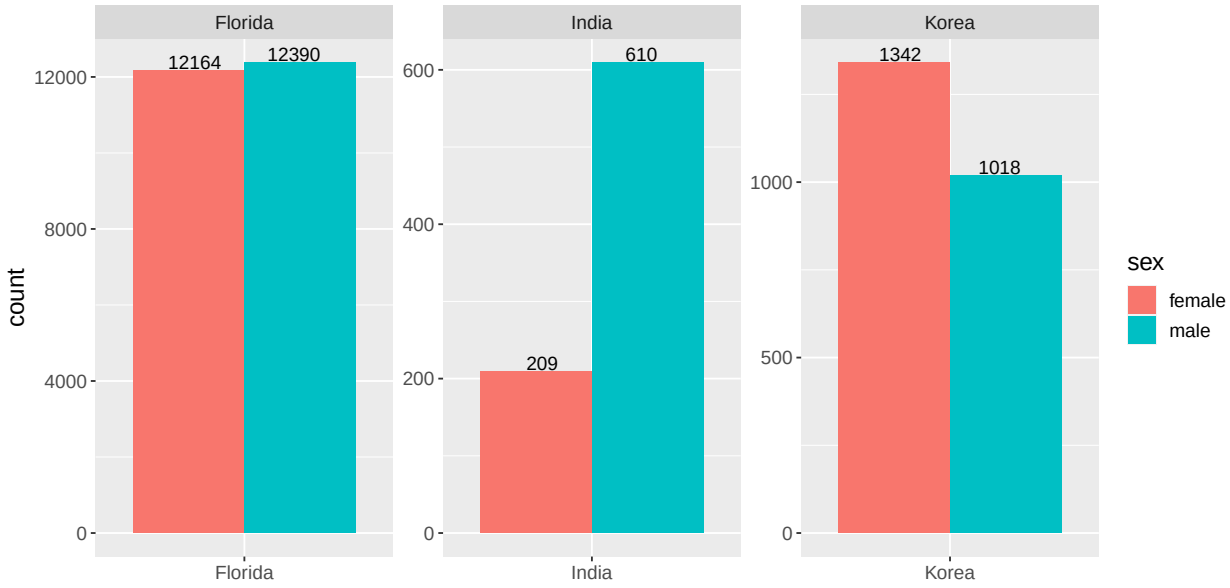

Figure 4: Breakdown of sex by country

Appendix A: Graphs and Tables

Table 1: Korea end status by sex

end.state	Deceased	Isolated	Released	Total
sex				
female	13 (1.0%)	948 (70.6%)	381 (28.4%)	1342 (100.0%)
male	29 (2.8%)	719 (70.6%)	270 (26.5%)	1018 (100.0%)
Total	42 (1.8%)	1667 (70.6%)	651 (27.6%)	2360 (100.0%)

Table 2: India end status by sex

end.state	Deceased	Hospitalized	Recovered	Total
gender				
female	7 (3.3%)	188 (90.0%)	14 (6.7%)	209 (100.0%)
male	16 (2.6%)	562 (92.1%)	32 (5.2%)	610 (100.0%)
Total	23 (2.8%)	750 (91.6%)	46 (5.6%)	819 (100.0%)

Table 3: Florida end status by sex

end.state	Alive	Dead	Total
sex			
female	11898 (97.8%)	266 (2.2%)	12164 (100.0%)
male	11943 (96.4%)	447 (3.6%)	12390 (100.0%)
Total	23841 (97.1%)	713 (2.9%)	24554 (100.0%)

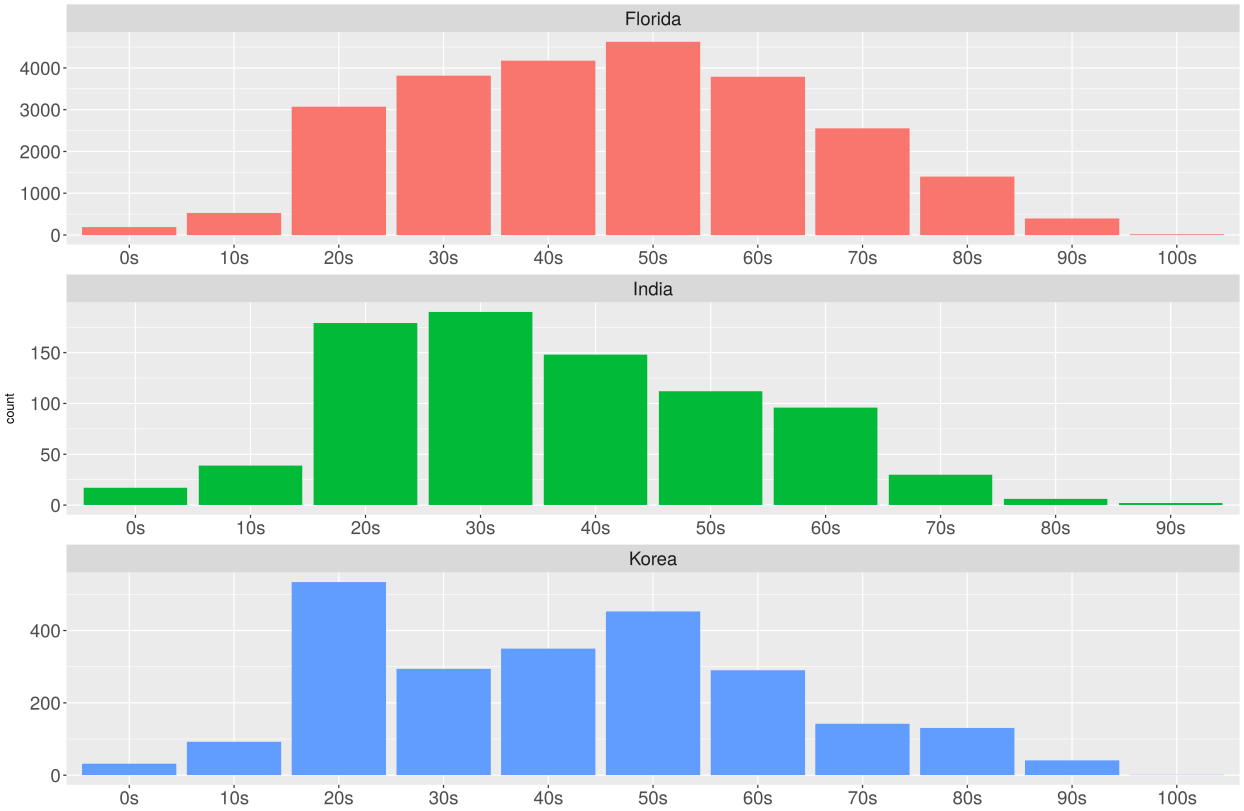

Figure 5: Breakdown of age by country

Table 4: Florida Age by End Status

end.state	Alive	Dead	Total
Age			
0s	190 (100.0%)	0 (0.0%)	190 (100.0%)
10s	529 (100.0%)	0 (0.0%)	529 (100.0%)
20s	3073 (100.0%)	1 (0.0%)	3074 (100.0%)
30s	3803 (99.7%)	12 (0.3%)	3815 (100.0%)
40s	4145 (99.4%)	25 (0.6%)	4170 (100.0%)
50s	4571 (98.9%)	52 (1.1%)	4623 (100.0%)
60s	3673 (97.0%)	114 (3.0%)	3787 (100.0%)
70s	2359 (92.4%)	195 (7.6%)	2554 (100.0%)
80s	1176 (84.1%)	222 (15.9%)	1398 (100.0%)
90s	308 (77.4%)	90 (22.6%)	398 (100.0%)
100s	14 (87.5%)	2 (12.5%)	16 (100.0%)
Total	23841 (97.1%)	713 (2.9%)	24554 (100.0%)

Table 5: Korea Age by End Status

end.state	Deceased	Isolated	Released	Total
Age				
0s	0 (0.0%)	24 (75.0%)	8 (25.0%)	32 (100.0%)
10s	0 (0.0%)	65 (70.7%)	27 (29.3%)	92 (100.0%)
20s	0 (0.0%)	348 (65.2%)	186 (34.8%)	534 (100.0%)
30s	1 (0.3%)	203 (69.0%)	90 (30.6%)	294 (100.0%)
40s	0 (0.0%)	236 (67.4%)	114 (32.6%)	350 (100.0%)
50s	3 (0.7%)	312 (68.9%)	138 (30.5%)	453 (100.0%)
60s	8 (2.8%)	228 (78.6%)	54 (18.6%)	290 (100.0%)
70s	12 (8.5%)	107 (75.4%)	23 (16.2%)	142 (100.0%)
80s	14 (10.7%)	107 (81.7%)	10 (7.6%)	131 (100.0%)
90s	4 (9.8%)	36 (87.8%)	1 (2.4%)	41 (100.0%)
100s	0 (0.0%)	1 (100.0%)	0 (0.0%)	1 (100.0%)
Total	42 (1.8%)	1667 (70.6%)	651 (27.6%)	2360 (100.0%)

Table 6: India Age by End Status

end.state	Deceased	Hospitalized	Recovered	Total
Age				
0s	0 (0.0%)	16 (94.1%)	1 (5.9%)	17 (100.0%)
10s	0 (0.0%)	36 (92.3%)	3 (7.7%)	39 (100.0%)
20s	0 (0.0%)	165 (92.2%)	14 (7.8%)	179 (100.0%)
30s	2 (1.1%)	182 (95.8%)	6 (3.2%)	190 (100.0%)
40s	3 (2.0%)	137 (92.6%)	8 (5.4%)	148 (100.0%)
50s	2 (1.8%)	106 (94.6%)	4 (3.6%)	112 (100.0%)
60s	10 (10.4%)	80 (83.3%)	6 (6.2%)	96 (100.0%)
70s	5 (16.7%)	24 (80.0%)	1 (3.3%)	30 (100.0%)
80s	1 (16.7%)	3 (50.0%)	2 (33.3%)	6 (100.0%)
90s	0 (0.0%)	1 (50.0%)	1 (50.0%)	2 (100.0%)
100s	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Total	23 (2.8%)	750 (91.6%)	46 (5.6%)	819 (100.0%)

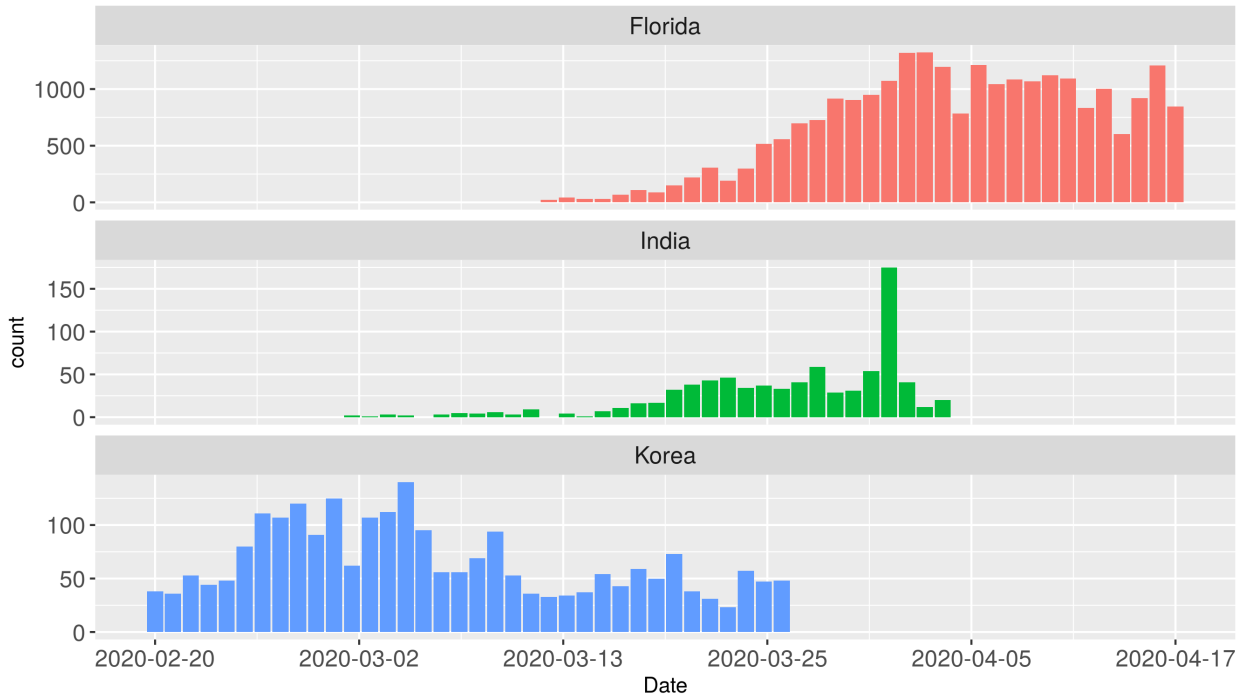

Figure 6: Daily cases for each Country

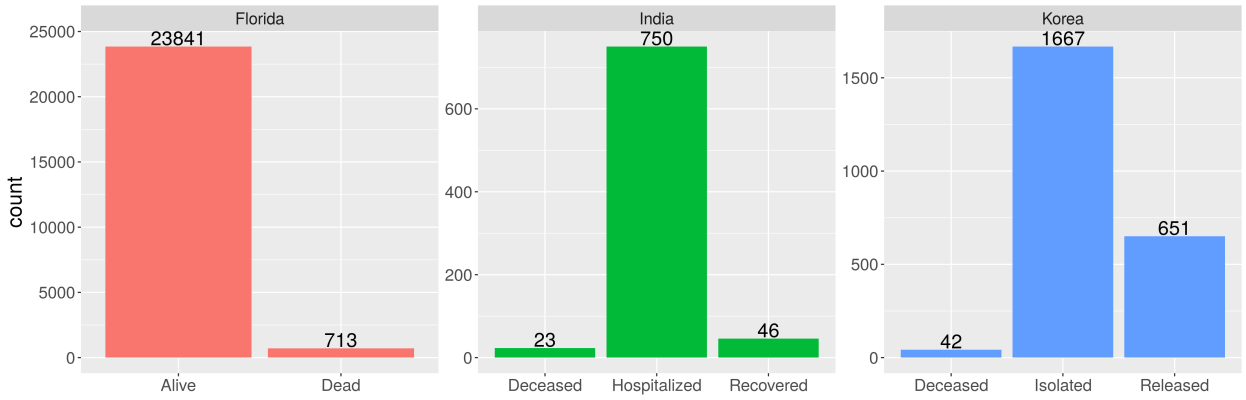

Figure 7: Patient status at the end of study period. Dead/Deceased indicates a non-censored observation.

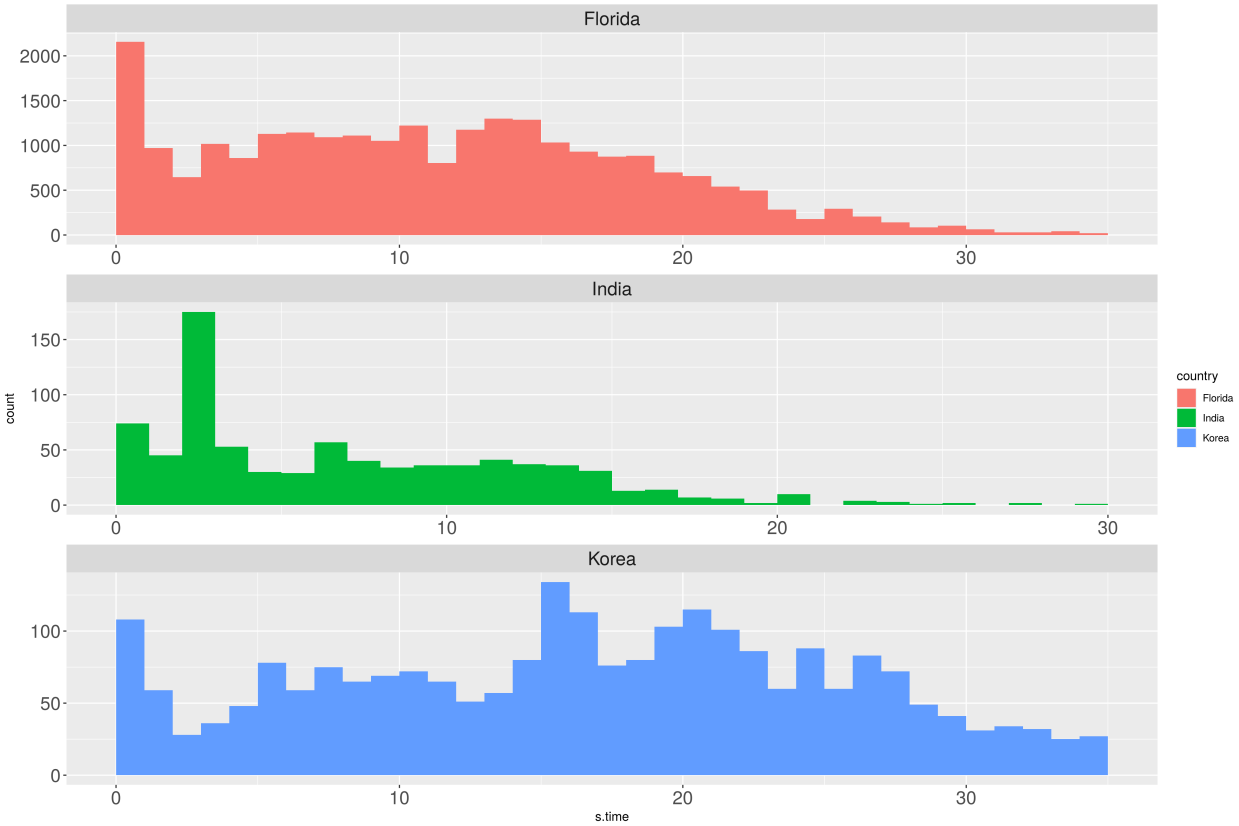

Figure 8: Distribution of survival times for each country.

	mean	Median	Std. Dev.
Florida	11.82	11	7.31
India	7.57	7	5.53
South Korea	17.08	17	8.87

Table 7: Simple desriptive statistics for survival times

```
survdiff(Surv(s.time, end.status)~country+strata(age.group, sex), data=combined)
```

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Florida	24554	713	709.03	0.022	0.257
India	819	23	7.18	34.864	35.730
Korea	2360	42	61.79	6.337	7.085

Table 8: Results For Stratified Log-Rank on Country. Chisq=42, df=2, p=8e-10

```
pairwise_survdiff(Surv(s.time, end.status)~country, data=combined)
```

	Florida	India
India	0.1436	
Korea	0.0000	0.0073

Table 9: P-values for pairwise comparison of Countries

```
survdiff(Surv(s.time, end.status)~sex+strata(age.group, country), data=combined)
```

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Female	13715	286	378	22.3	44.3
Male	14018	492	400	21.1	44.3

Table 10: Results For Stratified Log-Rank on Sex. Chisq=44.3, df=1, p=3e-11

```
survdiff(Surv(s.time, end.status)~age.group+strata(country, sex), data=combined)
```

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
<20	899	0	23.5	23.48	24.44
20s	3787	1	105.5	103.54	121.27
30s	4299	15	125.2	96.97	116.20
40s	4668	28	131.7	81.62	98.60
50s	5188	57	145.6	53.96	66.62
60s	4173	132	119.7	1.27	1.51
>70	4719	545	126.8	1378.83	1659.83

Table 11: Results For Stratified Log-Rank on age group. Chisq=1756, df=6, p=<2e-16

```
pairwise_survdiff(Surv(s.time, end.status)~age.group, data=combined)
```

	<20	>70	20s	30s	40s	50s
>70	0.0000					
20s	0.6338	0.0000				
30s	0.0880	0.0000	0.0014			
40s	0.0272	0.0000	0.0000	0.0880		
50s	0.0027	0.0000	0.0000	0.0000	0.0104	
60s	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 12: P-values for pairwise comparison of age.group

```
anova(mod.cox1, mod.cox2)
```

	loglik	Chisq	DF	P-value
Model 1	-6664.8	–	–	–
Model 2	-7349.1	1368.4	6	<.0001

Table 13: Partial log-likelihood test to determine whether age.group is significant
Model 1: ∼ age.group +sex + strata(country)
Model 2: ∼ + strata(country)

```
mod.cox3=coxph(Surv(s.time, end.status)~Age+sex+strata(country), data=combined)
cox.zph(mod.cox3)
```

	chisq	df	p
Age	0.444	1	0.510
sex	2.424	1	0.120
GLOBAL	3.011	2	0.220

Table 14: Results from PH test for stratified model with numeric age

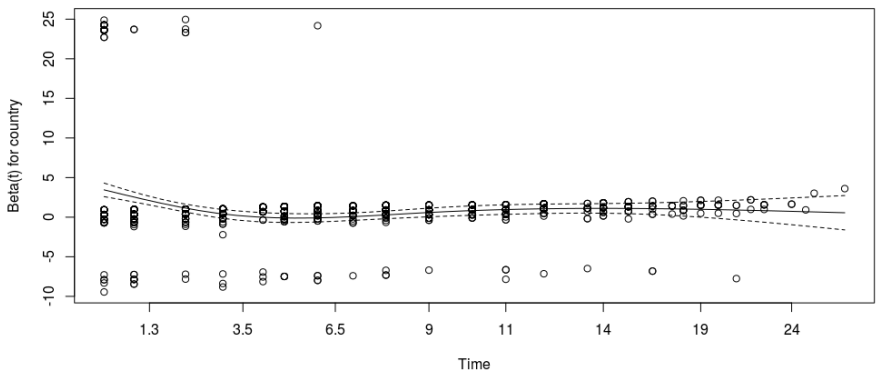

Figure 9: Shoenfeld residuals for country variable.

```
coxph(Surv(tstart, s.time, end.status) ~
        Age + sexmale + countryIndia + countryKorea + countryIndia:s.time +
        countryKorea:s.time + sexmale:s.time+Age:s.time, data=x.mat.Age)
sum.modF=summary(mod.extF)
```

	coef	exp(coef)	se(coef)	z	p
Age	0.086	1.089	0.004	20.860	<.0001
sexmale	0.404	1.497	0.123	3.288	0.0010
countryIndia	3.376	29.242	0.387	8.716	<.0001
countryKorea	0.183	1.201	0.258	0.711	0.4769
countryIndia:s.time	-1.022	0.360	0.289	-3.540	0.0004
countryKorea:s.time	-0.095	0.909	0.033	-2.841	0.0045
sexmale:s.time	0.024	1.024	0.014	1.639	0.1012
Age:s.time	0.000	1.000	0.000	0.474	0.6354

Table 15: Full extended Cox model. LRT=1667, 8 df, p<.0001, 775 events, n=354399

Appendix B: R code see "project code.R"