**ADVANCED DATA ANALYSIS**

**PROJECT**

**CONCEPT PROPOSAL INSTRUCTIONS**

**TITLE OF PROJECT:** Survival Analysis of COVID-19 in South Korea

**STUDENT(S):** Sam Hsu and Sherly Boddu

**BACKGROUND:**

Coronavirus Disease 2019 (COVID-19) was first reported in Wuhan city, Hubei Province, China in December, 2019.1 Since its emergence, the virus that causes COVID-19 has rapidly spread across six continents, and remains responsible for 375,000 confirmed cases and 16,370 deaths globally.2 Novel evidence suggests that the virus can result in mild symptoms to severe cause of fatality.3,4 Moreover, older adults and those with comorbidities (cardiovascular disease, diabetes, respiratory disease, and cancer) are at greater risk of death from COVID-19 infections.3-5 Hence, the aim of this article is to provide a timely review of the characteristics that impact survival of COVID-19.

More specifically, this paper will focus on cases identified in South Korea due to the availability of transparent data and the success of transmission prevention efforts.6 South Korea’s unique public health measures and testing capacity provides an accurate assessment of confirmed cases.5,6

**OBJECTIVES**:

1. Estimate probability of survival among confirmed COVID-19 cases in the dataset.
2. Determine demographic characteristics that contribute to survival from COVID-19
3. Identify environmental and regional characteristics that contribute to survival from COVID-19
4. Add to the body of COVID-19 literature to help guide surveillance and public health measures in mitigating the 2020 pandemic.

**APPROACH**:

We will use coronavirus (COVID-19) data provided by the Korean Centers for Disease Control and made publicly available through the online data science community Kaggle.7 This dataset currently includes every positively diagnosed case of COVID-19 occurring in South Korea until March 20, 2020. Since this dataset is live and continues to be updated with new cases, we will download new data until the point when we are writing our results.

Since we are using a survival analysis approach, we will use confirmation and release date/deceased date to count the number of days from diagnosis to resolution as the dependent variable. We will use age, sex, infection case (i.e., overseas inflow, direct contact, or church/hospital hub) as covariates. We will also consider regional characteristics like school ratio and elderly population ratio as well as weather data at time of infection like temperature, wind, and humidity as potential additional covariates, although these do not seem likely to influence survival.

As discussed previously, the dataset is publicly available on Kaggle. To ensure reproducibility of analysis, we plan to use RMarkdown to document analysis. We will have one team member write the analysis code and the other conduct a thorough check. Analysis will be conducted using R version 3.6.2. We will summarize our data using descriptive statistics like mean age of cases, proportion of men, and cluster concentrations by geography. For survival analysis, we will use the Cox proportional hazards regression to model time to death for different age categories.

**APPENDICES**:

*References*

1. Deng S-Q, Peng H-J. Characteristics of and Public Health Responses to the Coronavirus Disease 2019 Outbreak in China. *J Clin Med*. 2020;9(2). doi:10.3390/jcm9020575
2. Coronavirus: A timeline of the pandemic from China’s first case to now - Business Insider. https://www.businessinsider.com/coronavirus-pandemic-timeline-history-major-events-2020-3. Accessed March 25, 2020.
3. Ruan Q, Yang K, Wang W, Jiang L, Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. *Intensive Care Med*. March 2020:1-3. doi:10.1007/s00134-020-05991-x
4. Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. [The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China]. *Zhonghua Liu Xing Bing Xue Za Zhi*. 2020;41(2):145-151. doi:10.3760/cma.j.issn.0254-6450.2020.02.003
5. Lim J, Jeon S, Shin HY, et al. Case of the index patient who caused tertiary transmission of coronavirus disease 2019 in Korea: The application of lopinavir/ritonavir for the treatment of COVID-19 pneumonia monitored by quantitative RT-PCR. *J Korean Med Sci*. 2020;35(6). doi:10.3346/jkms.2020.35.e79
6. Commitment, transparency pay off as South Korea limits COVID-19 spread – EURACTIV.com. https://www.euractiv.com/section/coronavirus/news/commitment-transparency-pay-off-as-south-korea-limits-covid-19-spread/. Accessed March 25, 2020.
7. Korea Centers for Disease Control & Prevention. Kaggle. https://www.kaggle.com/kimjihoo/coronavirusdataset. Updated March 21, 2020. Accessed March 24, 2020.

*Project timeline*

3/25 - 4/1: Data cleaning and preparation of final analytic dataset.

4/1 - 4/8: Preparation of descriptive statistics table and initial survival analysis in R.

4/8 - 4/15: Verification of analysis results and interpretation of results.

4/15 - 4/22: Compilation of results into presentation.

*Roles*

Sam will conduct data cleaning and initial creation of the model in R. He will also be responsible for formatting the descriptives table. Sherly will support and verify the analysis and create figures for the presentation. She will also summarize relevant points for discussion. Both will be equally responsible for delivering the presentation.